Self-Directedness and Resoluteness.
The Two Dimensions of Autonomy

Dissertation

zur Erlangung des akademischen Grades

Doctor philosophiae

(Dr. phil.)

eingereicht

an der Philosophischen Fakultät I

der Humboldt-Universität zu Berlin

von M.A. Jan Prause-Stamm

Prof. Dr. Jan-Hendrik Olbertz, Präsident der Humboldt-Universität zu Berlin

Prof. Michael Seadle, PhD, Dekan der Philosophischen Fakultät I

Gutachter

1. Prof. Dr. Michael Pauen

2. Prof. Dr. Dr. Henrik Walter

3. Prof. Dr. Kirsten Meyer

Tag der Promotion (Disputation): 11. Februar 2013

Danksagung

Diese Arbeit ist im Rahmen des von der Volkswagen-Stiftung geförderten Projekts „Autonomie – Handlungsspielräume des Selbst“ entstanden. Ich bedanke mich bei der Volkswagen-Stiftung für das Stipendium, das ich zur Anfertigung meiner Dissertation erhalten habe. Besonderer Dank gilt meinen beiden Betreuern, Prof. Dr. Michael Pauen und Prof. Dr. Dr. Henrik Walter, die mich stets unterstützt und bestärkt haben. Schließlich möchte ich mich auch bei der *Berlin School of Mind and Brain* bedanken, deren Graduiertenprogramm ich während meiner Promotionszeit absolvieren durfte.

Dortmund, 22.06.2012

Jan Prause-Stamm

Summary

I explore and explicate a notion of personal autonomy which has its sources (1) in the political concept of autonomy as it was developed in ancient Greece, (2) in Kant's theory of autonomy as a property of person, and (3) in Frankfurt's individualistic approach towards autonomy. From a systematic point of view I conceptualize autonomy as a natural and gradual property of persons, which is not tied to norms, and which differs from free will. Autonomy deals with an agent's authentic expression of her standpoint and her aptitude in prevailing in conflicts or difficult situations. The central insight of this study is that resolute agency is an essential aspect of autonomy. Dispositional autonomy is constituted by an agent's dispositions and abilities to overcome obstacles. The autonomy debate underemphasized this aspect of autonomy by solely focusing on self-directed agency. But as important as self-directedness is for autonomous agency, without resoluteness, an agent falls short of being autonomous. The personal strength that resoluteness consists in is a hallmark of autonomous agents.

Zusammenfassung

Diese Arbeit untersucht und expliziert ein Verständnis von Autonomie, das seine Wurzeln (1) in der antiken politischen Autonomiekonzeption, (2) in Kants Theorie von Autonomie als Eigenschaft von Personen und (3) in Frankfurts individualistischem Verständnis von Autonomie hat. Autonomie wird systematisch aufgefasst als eine natürliche und graduelle Eigenschaft von Personen, die nicht an spezifische Normen gebunden ist und sich vom freien Willen abgrenzen lässt. Für Autonomie ist es wesentlich, dass die Person ihren eigenen, authentischen Standpunkt ausdrückt und dazu fähig und disponiert ist, Hindernisse zu überwinden und sich in Konflikten zu behaupten. Die zeitgenössische Autonomiedebatte hat diesem Aspekt von Autonomie zu wenig Beachtung geschenkt und sich nahezu ausschließlich auf Selbstgerichtetheit konzentriert. Ich argumentiere, dass Selbstgerichtetheit eine wesentliche, aber nicht hinreichende Dimension von Autonomie ist. Eine vollständige Konzeption von Autonomie muss die zweite Dimension von Autonomie, nämlich Entschlossenheit, mitberücksichtigen. Die persönliche Stärke, die Entschlossenheit auszeichnet, ist ein wesentliches Merkmal autonomer Akteure.

Contents

Introduction

The beginnings of autonomy are marked by tragedy and heroic death. Antigone, the very first figure that we know of that has ever been called autonomous, defies the laws of Thebes and the commands of its ruler Creon – and follows her own judgment. By doing what her conscience tells her to do rather than bowing down to the demands of those in power, Antigone condemns herself to death. But it is her choice and her choice alone. Listen to what the chorus sings about Antigone:

> "under your own law, alive, alone and unique
> of mortals, you will descend to Hades."[1]

The Chorus sings about Antigone who is going to Hades, which is a metaphorical way to say that she is condemned to die. Her case is special because according to the Chorus, she is the first person that ever had to die "under her own law." Sophocles uses the term 'autonomy' to describe Antigone's attitude or character in defying her uncle and the laws of Thebes. His drama is the oldest document we know in which the notion of autonomy is used.[2]

What was her deed? Antigone buried her brother Polynices who had, according to Thebes' laws, forfeited his rights to a proper burial when he turned traitor and attacked the city. Creon's orders left no room for doubt: Polynices' corpse had to be left outside of the city as prey for the animals. Disobedience would be punished with death. Antigone knew this. But she did not let herself be intimidated into submission. Instead, she did what she judged to be right although she knew that she would have to pay with her life. She brought this death sentence upon herself by acting under her own law, that is, by acting autonomously.

---

[1] Sophocles: *Antigone*, trans. by William Blake Tyrrell/Larry J. Bennett (http://www.stoa.org/diotima/anthology/ant/antigstruct.htm), 821. Also compare 875.

[2] Sophocles' Antigone contains "[t]he earliest datable occurrence of the adjective (441 B.C.)." Martin Ostwald (1982): *Autonomia: Its Genesis and Early History* (Atlanta: Scholars Press), 10. Note, however, that Sophocles uses the notion of autonomy in a metaphorical way when he refers to Antigone, an individual person and not a *polis*, as autonomous. Ostwald points out that the concept of autonomy was in Sophocles' time systematically established only as a political notion and not as a notion which refers to properties of persons. Sophocles took the notion of autonomy out of the political discourse, where it referred to a status of *poleis'*, and applied it metaphorically to a person. As I discuss in Chapter 1, it was most notably Immanuel Kant who developed a systematical understanding of autonomy as a property of persons.

Antigone exemplifies a very important concept of autonomy. The aim of this study is to explore and explicate this concept in greater detail. To start with, let me mention some more examples of the phenomenon I have in mind. Martin Luther is another good illustration for this sense of autonomy. He famously declared, "Here I stand. I can do no other."[3] With these words Luther answered a church tribunal that asked him whether he still believed in what he had written about the church and the pope. The tribunal expected Luther to recant. It was clear that he would suffer dearly if he were to uphold his position. When confronted with the question whether or not he still stood behind what he had said, Luther gave his famous answer.[4]

Although these two examples conjure up the images of singular acts, the concept of autonomy that I am after is particularly concerned with an agent's personal traits – her character – and how they find expression in her whole way of life. Autonomous people shape their lives according to their own desires, beliefs, and values. Think of someone like Marie Curie, for example. She was born in a society that discouraged women from becoming scientists and prevented them from studying at a university. She circumvented this obstacle by moving from Poland to France. There women were at least allowed to study. She encountered a scientific community that was dominated by men and that did not welcome female scientists. Nevertheless, she managed to become one of the most important scientists of the 20th century. Decorated with two Nobel prizes, she was the first woman to become a professor at the Sorbonne. We can only imagine how difficult it was for her to achieve this kind of academic success. However, she mustered the inner strength and courage to prevail against social and cultural obstacles in order to do what she wanted to do. She lived her life in accordance with her own values and desires and did not conform to the

[3] It has been disputed that Luther actually uttered these exact words. This, however, does not render the aim of the example mute, namely to illustrate a certain concept of autonomy. If you are in doubt that the example is historically valid, just imagine an agent who actually stands his ground in front of a church tribunal because he thinks that this is the right thing to do. Compare for a historical discussion: Roland H. Bainton (1950): *Here I Stand. A Life of Martin Luther* (Nashville: Abingdon Press).

[4] Apart from the question whether or not this example is historically true – a question which is irrelevant for the purposes of our discussion (compare FN 3) – some people might reject the example because they interpret Luther's actions not so much as being autonomously guided, but as being the upshot of God's commands. After all, isn't it true that Luther thought of himself only as a humble servant of God and thus only acted as a spokesman of God? In answer to this, let me emphasize again, that I do not want to engage in a historical debate. I shall simply interpret Luther's actions as expressing his own authentic standpoint. I view him as a person who deeply cared about his religious conscience and his spirituality and acted accordingly. From this perspective he was not just a tool of God but an agent who strongly fought for his own convictions.

picture her society saw fit for women. Because of this, she is a very good example of an exceptionally autonomous person.

Or think of a great artist like Marcel Duchamp. He defied expectations and developed his art – and art in general – to ever new levels. Of course, he was never threatened with death or imprisonment, like Antigone or Luther. And it was probably harder for Marie Curie to navigate her way through a world that was antagonistic to the very core of how she envisioned her life than it was for Duchamp to explore blank spaces on the artistic landscape. However, what they have in common is that each of them managed to live up to their own judgments about what to do and how to live rather than leading a life shaped by social expectations and the demands of others.

All these examples point in the direction of a particular concept of autonomy. According to this concept, autonomy is attributed to persons who can find their own standpoint and make it manifest in their life. Autonomy combines the two ideas of having a standpoint and expressing it. The person who has a standpoint but betrays it fails to be autonomous, as does the person who does not find her own standpoint in the first place. Moreover, what the examples demonstrate and what is implied by the metaphor of making one's own standpoint manifest, is that autonomy comes at a price. You have to be prepared to fight for your own perspective in order to count as autonomous. This antagonistic aspect of autonomy is often neglected.[5] I argue that we need to bring it back to the foreground in order to adequately understand the concept of autonomy that is exemplified by agents such as Antigone and Marie Curie.

I just stated that that which characterizes the autonomous person is that she develops her own standpoint and expresses it in her life. She shapes her life in accordance with her own desires, beliefs, and values. These metaphorical characterizations of autonomy are certainly in need of clarification. What I have outlined so far describes the intuitive starting point for the following discussion. The notion of autonomy is not frequently used in everyday language. And as we will see briefly, its use in philosophy is very diverse. It is fair to say that there is not only one concept of autonomy, but many. Hence, an initial requirement for a study of

[5] Compare for example John Benson's characterization of the autonomous man: "The autonomous man has a mind of his own and a will of his own. He exercises independence in his thinking and in his decisions about practical affairs." John Benson (1983): 'Who is the Autonomous Man?', in: *Philosophy* (223), 5-17, 6. This characterization captures something quite essential about autonomy. It remains mute, however, about the antagonistic element of autonomy.

autonomy is to sketch the phenomenon that it is supposed to be about. I view it as one of the requirements for an adequate account of the concept of autonomy that I have in mind that it is able to explain in virtue of what people, like Antigone, Martin Luther, Marie Curie and Marcel Duchamp, are autonomous.

To get an idea about the diversity that characterizes the use of the notion of autonomy, look at this list by Gerald Dworkin:

> "It [the notion of autonomy] is used sometimes as an equivalent of liberty (positive or negative in Berlin's terminology), sometimes as an equivalent to self-rule or sovereignty, sometimes as identical with freedom of the will. It is equated with dignity, integrity, individuality, independence, responsibility, and self-knowledge. It is identified with qualities of self-assertion, with critical reflection, with freedom from obligation, with absence of external causation, with knowledge of one's own interests. It is equated by some economists with the impossibility of interpersonal comparisons. It is related to actions, to beliefs, to reasons for acting, to rules, to the will of other persons, to thoughts, and to principles. About the only features held constant from one author to another are that autonomy is a feature of persons and that it is a desirable quality to have."[6]

This plethora of different meanings, connotations, and systematical contexts makes it necessary to zoom in in much more detail on what is meant by autonomy in a certain context. The aim of my discussion is to explicate a particular concept of autonomy more systematically. It is the concept that is, among other things, exemplified by Antigone, Martin Luther, Marie Curie, or Marcel Duchamp. This concept is, as Joel Feinberg puts it, about "the capacity to govern oneself […] or the actual condition of self-government." Feinberg distinguishes four concepts of autonomy:

> "When applied to individuals the word 'autonomy' has four closely related meanings. It can refer either to the capacity to govern oneself, which of course is a matter of degree; or to the actual condition of self-government and its

[6] Gerald Dworkin (1988): *The Theory and Practice of Autonomy* (Cambridge: Cambridge University Press), 6.

> associated virtues; or to an ideal of character derived from that conception; or (on the analogy to a political state) to the sovereign authority to govern oneself, which is absolute within one's own moral boundaries (one's 'territory,' 'realm,' 'sphere,' or 'domain')."[7]

As will become clearer throughout the analysis, I am investigating a concept of autonomy that views autonomy as a dispositional property of persons and its actualization.[8]

In order to understand a certain phenomenon better, it is often helpful to look at cases where it gets lost or is absent. So far I have presented positive examples of autonomous agency. To further clarify the phenomenon I am interested in, I also want to highlight some paradigmatic examples of a lack or breakdown of autonomy. On a very general level, I conceive of conformist behavior as the opposite of autonomy. The conformist agent is characterized by just doing what others do without developing her own standpoint or by disregarding her own standpoint. I understand autonomy essentially as something that is contradicted by conformism. Let me introduce three other important cases of non-autonomy, namely compulsion, coercion, and manipulation. For all of these cases, it is true that they violate the kind of autonomy I explore in this study. If someone doubts whether autonomy is damaged or destroyed by compulsion, coercion, and manipulation, I can only note that this person appears to have a different concept in mind. I address a concept of autonomy according to which autonomy is, among other things, that which is violated by compulsion, coercion, and manipulation.[9]

---

[7] Joel Feinberg (1989): 'Autonomy', in: John Christman (ed.) (1989): *The Inner Citadel. Essays on Individual Autonomy* (Oxford: Oxford University Press), 27-53, 28.

[8] For a more explorative distinction of different meanings or dimensions of autonomy compare Rainer Forst (2005): 'Political Liberty: Integrating Five Conceptions of Autonomy', in: John Christman/Joel Anderson (eds.) (2005): *Autonomy and the Challenges to Liberalism* (Cambridge: Cambridge University Press), 226-242.

[9] Some philosophers conceptualize autonomy in a way that does not view conformism and autonomy as contradictory. Compare for example Richard Dworkin (1988) and Laura Waddell Ekstrom (2005): 'Autonomy and Personal Integration', in: James Stacey Taylor (ed.) (2005): *Personal Autonomy. New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy* (Cambridge: Cambridge University Press), 143-161. The basic idea is that autonomy is content-neutral. In contrast to that, I view autonomy as involving some substantial constraints. In particular, autonomy is contradicted by conformism and subservience. Marina Oshana argues in a similar spirit that "some states of affairs and some social roles are antithetical to autonomy." Marina A. L. Oshana (2002): 'The Misguided Marriage of Responsibility and Autonomy', in: *The Journal of Ethics* (6), 261-280, 274. My

A compulsive agent cannot help thinking or acting in a certain way. A compulsive handwasher, for example, washes her hands repeatedly without being able to refrain from it. Very severe addictions are a standard example of compulsion. Someone with a severe drug addiction compulsively takes the drug. She cannot resist it. This does not exclude the possibility that she makes very elaborate plans to get the drug and to organize her life in an otherwise more or less normal way (although most severe addicts are on a slippery slope downwards). What she lacks is the ability to refrain from taking the drug. An important feature of compulsion is that it circumvents the agent's rational capacities. That is, the agent is unable to let go of the compulsive behavior even if she is convinced that she has best reason to refrain from it. It might be the case that the agent has a reason for performing the compulsive action. This reason, however, does not explain why she is performing the behavior. Compulsive action is often treated as the standard case of action that is non-autonomous.[10]

A second paradigmatic case of non-autonomy is coercion. A person who is coerced to do something is acting non-autonomously. Coercion is a demand that is accompanied by the threat of severe negative consequences in the case of non-compliance. Coercion attempts to create a new and decisive reason for action by introducing a new consequence in the case of omission. This negative consequence is contingent on the actions of the coercer. Without the influence of the coercer this negative consequence would not occur. This contingency on the ill will of another person distinguishes coercion from comparable negative consequences of acting in particular ways. An example of coercion is a street robbery. Someone who threatens to shoot you if you do not relinquish your purse thereby creates a new reason for you to hand over your purse. The negative consequence for keeping your money is being shot. That this is a probable consequence depends on the robber's malicious intentions. If you act on this reason, you act under coercion. The concept of autonomy I want to explain takes this to be a paradigmatic example of non-autonomous action.

---

own perspective is also influenced by Daniel Friedrich (2011): *Autonomy and Subservience* (Manuscript).

[10] Compare for a different approach towards the autonomy of compulsive agents Sarah Buss (1994): 'Autonomy Reconsidered', in: *Midwest Studies in Philosophy* (19), 95-121. Buss argues that the "unwilling addict" might autonomously take the drug: "Indeed, far from getting off the hook, the unwilling addict's genuine *disapproval* of his addiction reflects the very *belief* that makes it possible for him to act autonomously." Sarah Buss (1994), 102. (Emphasize in the original. If not otherwise stated all emphasizes are in the original text.) Buss clearly has a different concept of autonomy in mind.

Manipulation is a third case of violating autonomy. As I understand it, manipulation is an influence that aims at bringing about a certain mental state in an agent, say a desire, belief, or intention by circumventing the agent's rational capacities. The manipulation functions either by circumventing an agent's rational capacities completely (for example, by influencing the agent on a purely emotional level), or the rational capacities of the agent are manipulated by giving false or incomplete information with the consequence that the agent has a false understanding of the situation. Other situations in which an agent receives false or incomplete information are different from manipulation in being not intentionally brought about with the aim of bringing the agent to believe something. The ultimate aim of manipulation is always practical. The manipulator wants the manipulated agent to behave in a specific way. Even if the direct consequence of a manipulation is the formation of a certain belief, the final goal is that this belief causes the agent to act in a certain way or to refrain from a certain action.

Alfred Mele gives a quite elaborate example of manipulation in which Beth, a talented but not very committed philosopher who values other things a lot more than doing philosophy, is thoroughly brainwashed. As a consequence, she becomes like Ann who cares more about philosophy than anything else and works 12 hours per day.

> "Beth is now, in the relevant sense, a 'psychological twin' of Ann. She is an industrious philosopher who thoroughly enjoys and highly values her philosophical work. Indeed, it turns out – largely as a result of Beth's new hierarchy of values – that whatever upshot Ann's critical reflection about her own values and priorities would have, the same is true of critical reflection by Beth. Her critical reflection, like Ann's, fully supports her new life style. […] When she carefully reflects on her preferences and values, Beth finds that they fully support a life dedicated to philosophical work, and she wholeheartedly embraces such a life and the collection of values that support it."[11]

[11] Alfred R. Mele (1995): *Autonomous Agents. From Self-Control to Autonomy* (Oxford: Oxford University Press), 145.

Mele refers to the concept under consideration when he points out that "it is difficult not to see her now, in light of all this, as heteronomous to a significant degree."[12] Hence, when I explicate the concept of autonomy, I aim at an understanding that can make sense of this judgment.

This concludes my initial characterization of the concept of autonomy under consideration. My aim is to explicate this concept that, as far as I see it, has been in the focus of much of the recent autonomy debate. At the same time, I am aware that, as Dworkin points out, a lot of different ideas are intermingled in this debate. I investigate one thread of it. In Chapter 1, I go on a historical journey and investigate three major steps in the history of the notion of autonomy. Chapter 2 takes up these thoughts and spells them out more systematically. At the end of Chapter 2, I will have delineated the concept that I investigate in this study. A central idea is that autonomy is a natural property and thus needs to be accounted for within a naturalistic framework. Chapter 3 is devoted to a discussion about what I regard as the hallmarks of a naturalistic account of autonomous agency. Chapter 4 provides the action theoretical background for my understanding of autonomy. I analyze the notion of agential control and argue that actions are exercises of control. I discuss how our agential control is realized within a naturalistic framework and emphasize the importance of intentions in this context. Now, the autonomous agent is not just able to act, but can make use of this ability in specific ways. She has her own authentic standpoint which determines what she does. Whereas Chapter 4 deals with the question how we translate our standpoint into action, that is, how we can shape our lives, Chapter 5 explores the standpoint metaphor. The focus of this chapter is the question of how to systematically account for the self vs. non-self distinction. In Chapter 6, I develop some thoughts that are meant to support the idea that the concept of autonomy under investigation is not sufficiently captured in terms of self-directedness. The upshot of this discussion is that in addition to self-directed agency, autonomy is also characterized by what I want to call resolute agency. Finally, Chapter 7 explicates in more detail what I mean by resolute agency and how this constitutes autonomy.

[12] Alfred R. Mele (1995), 145.

1. The Notion of Autonomy – A Three-Step History

The notion of autonomy has a diverse and multifaceted history that gave rise to a variety of different understandings of what autonomy is. Different usages of the notion of autonomy evolved out of the same historical development. In this chapter, I highlight the three steps in this development that I take to be the most important ones with respect to the concept of autonomy that I want to explore. This chapter will pave the way for delineating the phenomenon that forms the focus of this study of autonomy. Against this historical background, I will take up the discussion in Chapter 2 in a more systematical manner.

The three milestones in the history of the notion of autonomy are the emergence of the idea of autonomy as a distinct political notion in the fifth century BC in Greece, Immanuel Kant's transference of the idea of autonomy to the personal level, and finally the move towards an individualistic sense of autonomy by Harry Frankfurt. I discuss each of these steps in this chapter. In section 1.1, I deal with the origins of the notion of autonomy in ancient Greece. An essential characteristic of this understanding is that autonomy is an *antagonistic* notion, that is, it deals with real or possible conflicts and, in particular, with the aptitude of prevailing in such conflicts. This focus on strength and assertiveness is a key aspect of my understanding of autonomy. In section 1.2, I discuss how Kant turned autonomy into a *personal* concept. Kant took the political idea of autonomy and applied it to persons. This step towards the personal opened the route for our modern understanding of personal autonomy. Finally, in section 1.3, I present Harry Frankfurt's approach towards autonomy, according to which autonomy is grounded in the *individual characteristics* of a person. Frankfurt is a starting point for the recent debate on autonomy.

1.1 Ancient Greece – The Political Origins

The notion of autonomy has its roots in ancient Greece. Its birth as a distinctive concept lies somewhere in the middle of the fifth century BC. Our best source for shedding light on the origins of the notion of autonomy is Thucydides' history of the Peloponnesian War.[13] Thucydides sets out to give us an exhaustive picture of the circumstances and historical developments that finally lead to the Peleponnesian War

[13] Thucydides (2009): *The Peloponnesian War*, trans. by Martin Hammond (Oxford: Oxford University Press).

between the two major powers in Greece: Sparta and Athens. From Thucydides' historical account about the political situation in Greece, we can extract, at least in outline, how the notion of autonomy became a distinct concept in the political discourse at that time. The following reconstruction of the original meaning of autonomy is indebted to the seminal work of the two classicists Markus Ostwald and Thomas Figueira on the origins of the notion of autonomy.[14]

The notion of autonomy evolved in a context of political struggles. The Delian League, which had been founded by Greek *poleis* as an alliance against Persia, was in disarray. Some *poleis* felt that Athens, the leader of the Delian League, had begun to exploit its status as the most powerful member and leader of the alliance. It demanded, for example, that the other members of the League pay a tribute to Athens. What had started out as an enterprise between more or less equal partners was slowly turning into an instrument of Athenian ambition. In the face of a real or imminent loss, some *poleis* used the rhetoric of autonomy in order to invoke their traditional rights. The city states started to use 'autonomy' as a distinct concept because a political independence that they had enjoyed previously came under external pressure. In particular, weaker members of the Delian League felt threatened by the increasing Athenian dominance. They feared that Athens would impinge upon their political integrity. As a reaction, they demanded to maintain their traditional political and military rights. A conflict like this finally led to the outburst of the Peloponnesian war. Aigina felt violated in its autonomy by Athens. It asked Sparta for help, arguing that Athens was not respecting Aiginetian autonomy. Sparta readily supported Aigina's claim, and in due course, the Peloponnesian War broke out.

To our modern ears, autonomy sometimes evokes connotations of autarky, absolute control, or total independence from influences. It is important to notice that the notion of autonomy had, in its beginnings, a rather narrow meaning compared to these modern intuitions. The original notion of autonomy meant only a restricted independence from outside interferences: "a form of αύτονομία can be envisaged which does involve the payment of tribute and which does affect the administration of justice."[15] A *polis* could be autonomous although it was liable to pay tribute and was not independent of external influence in its justice system. Autonomy was not an all-

[14] Compare Markus Ostwald (1982) and Thomas Figueira (1990): 'Autonomoi Kata Tas Spondas (Thucydides 1.67.2)', in: *Bulletin of the Institute of Classical Studies* (37), 63-88.
[15] Markus Ostwald (1982), 7.

encompassing right to be sovereign in all matters of the *polis*, but was limited to certain domains. Another tension to some modern intuitions concerning autonomy is that, at least in some usages, autonomy was understood to be something that was conceded by a superior power to a weaker *polis*. In this understanding, autonomy was not understood as a natural right, but as a political concession. One of the consequences was that the superior power itself would not proclaim to be autonomous because this would imply that it was dependent in some ways on others: "αύτονομία is not something which major powers claim for themselves but it is a concession or a recognition which they extend to states less powerful than themselves."[16]

The limits to one's sphere of self-determination, as well as the dependency of one's autonomy on the concession of a superior power, violate such absolute approaches towards autonomy like that of Kant or, more recently, Susan Wolf,[17] who argue that any form of dependency or external interference nullifies autonomy. For the Greeks of the fifth century, autonomy was not only compatible with some forms of dependency and interference. The concept gained its meaning in the first place as referring to a state's sphere of independence within a context of dependencies. In other words, autonomy was a status that could only apply to political units that were embedded in a net of dependencies. Total independence was not the imagined ideal of autonomy.

According to Ostwald's analysis of agreements and contracts between *poleis*, the political status that people referred to with the notion of autonomy was rather the norm. A *polis* was justified by tradition to retain its autonomy: "αύτονομία is regarded as the normal traditional status to which the states to whom it is applied are believed to be historically entitled." And once this norm came under pressure, people started to frame their demands in terms of autonomy: "the contingent independence expressed by αύτονομία was not conceptualized until the independence of some states came to be challenged."[18] In this sense, autonomy is conceived of as a right. And it is brought into the discourse because of alleged infringements of this right.

We have already seen that autonomy meant the status of a political unit that had to deal with opposing or interfering powers. According to Ostwald, this

[16] Markus Ostwald (1982), 7 f.
[17] Susan Wolf (1990): *Freedom Within Reason* (Oxford: Oxford University Press).
[18] Markus Ostwald (1982), 14 f.

understanding of autonomy as a status that was held by a polis in the context of possible conflicts became a central aspect of its meaning. Ostwald's understanding heavily relies on E.J. Bikerman's work: "It is a great merit of his [Bikerman's] work to have established that αύτονομία differs from έλευθερία in being a concept in interstate relations, in that the independence of the 'autonomous' state stands always in the shadow of a stronger power."[19] In his own investigation, Ostwald corroborates this point and emphasizes that, once the notion of autonomy was used in a distinct sense, it described the status of a political unit that is confronted with a stronger power. Autonomy was basically a concept that described how a weaker *polis* is related to a stronger one:

> "it [the notion of autonomy] is always used of a weaker state which tries to assert its independence in the face of a major power, but never the independence of the stronger power, our thesis is that αύτονομία developed in an attempt by weaker states to find constraints with which to inhibit the exercise of power by a stronger state over them."[20]

According to this interpretation, the primary use of the notion of autonomy has been a defensive one. This contrasts with one of the modern concepts of autonomy that we encountered in Feinberg's distinction, namely autonomy as an ideal that everybody should adhere to.

This understanding is also supported by Figueira's interpretation, although Figueira disagrees with Ostwald's claim that there was one underlying meaning of the notion of autonomy common to all usages. According to Figueira, the Spartan and the Athenian understanding of autonomy were slightly different. In his reading, Sparta understood autonomy as being more or less identical to "freedom" and "independence," although it is not entirely clear how Figueira understands these notions in this context. It is *prima facie* plausible to assume that he wants to denote a more robust right to be free of external interferences when he places the Spartan use of 'autonomy' near the notions of freedom and independence. However, this reading

[19] Markus Ostwald (1982), 1.
[20] Markus Ostwald (1982), 1.

appears not to be adequate because Figueira concedes that the Spartan notion of autonomy "is also used of poleis which seem to external appearances to have accepted the hegemony of others and to have been thereby constrained in their foreign policy, if not in their internal political life."[21] On the one hand, Figueira appears to view Sparta's notion of autonomy as meaning some sort of unrestricted independence. On the other hand, this status of autonomy as independence was compatible with bowing one's head to a hegemonic power: "The Spartan definition of autonomy appears to be close to the sense of 'independence' just noted. Autonomy, however, does not preclude acknowledgment of Sparta as *hēgemōn*."[22]

The Athenian notion of autonomy was, in Figueira's reading, primarily about the status of allies: "In the terminology of Thucydides on the Athenian Empire, autonomy denotes the status of allied cities maintaining an independent military establishment, who were thereby exempt from the mechanism for exaction of tribute."[23] This idea is akin to the description of autonomy as a restricted independence. Those allies did not possess complete independence. They were expected to follow Athens' lead in times of war. Indeed, it was one of the hallmarks of autonomy that a polis could provision ships in the case of war: "the distinction between autonomy and subject status is sharper and specifically marked by the provision of military forces to the Athenian alliance and freedom from the assessment of tribute."[24] Figueira goes on by pointing out that "[t]he association between the provision of ships and autonomy is so close" that *poleis* were sometimes counted "as autonomous by virtue of provision […] of ships."[25]

That military power was an integral aspect of autonomy can be seen not only with respect to the possession of a fleet, but also the possession of fortifications. "Another token of autonomy was the possession of walls […], so that their demolition and the surrender of ships were interpreted together as a demotion in status […] The possession of ships and fortifications amounts to the military capability which was an essential aspect of autonomy."[26] What this shows is that autonomy, in this sense at least, was not only understood as the right to have a certain sphere of sovereignty or

[21] Thomas Figueira (1990), 64.
[22] Thomas Figueira (1990), 64.
[23] Thomas Figueira (1990), 67.
[24] Thomas Figueira (1990), 67.
[25] Thomas Figueira (1990), 68.
[26] Thomas Figueira (1990), 68.

as the absence of external interferences. It also meant the capacity to engage in military conflicts and to defend oneself. This is especially noteworthy because there is a tendency in the contemporary debate to view autonomy as an internal matter and to neglect the relational aspects of autonomy. This is, at least compared to the original understanding of autonomy, too narrow a perspective. By relational aspects I mean that autonomy is concerned with how one relates to others. Are others dominating me, for example, or do I possess the strength to reject their demands? For the Greeks, a certain amount of strength was a constitutive part of autonomy. Figueira leaves no room for doubt about this: "There can be no pretense of autonomy in a context of total dependence on another for one's security."[27]

Here is another remark by Figueira that supports the idea that autonomy was originally understood to be a status that also described how a polis relates to other poleis:

> "Autonomy may have begun by meaning of independence of a city's internal decision-making process. The usage of the term in its Athenian connotation, however, focuses on liability or non-liability to pay tribute, on the possession of a fleet, and on the existence of fortifications. Nonetheless, it is mistaken to see in these characteristics only the outward trappings of internal autonomy, for they are essential aspects of autonomy itself. Clearly, the idea that one could have total internal independence without a military apparatus was inconceivable."[28]

We can learn from Thucydides that autonomy was a political notion that emerged in a context of potential infringements of a *polis'*s political integrity. *Poleis* that possessed military power and managed to uphold a sphere of self-determination, even though they had to deal with superior powers, were the paradigm cases of autonomy. This descriptive sense of the status of autonomy was accompanied by a normative one, according to which *poleis* had the right to be autonomous. Hence, in the beginning, autonomy was a political notion that referred to the *right* and the *power* of *poleis* to decide for themselves, at least to a certain extent, how to organize their own political and military affairs. Thucydides emphasizes that no *polis* could be autonomous if it

[27] Thomas Figueira (1990), 70.
[28] Thomas Figueira (1990), 88.

had no fleet and no fortifications. These military means were an integral part of a state's autonomy. The understanding of autonomy as an aptitude to prevail in conflicts is of particular importance for the concept of autonomy that I am exploring. I highlight it especially in Chapters 6 and 7.

1.2 Immanuel Kant – From the Political to the Personal

Autonomy started out as a political concept that applied to states or state-like entities.[29] The most important figure for transferring the notion of autonomy from the political to the personal level was Immanuel Kant. That it is natural for us to view autonomy as being concerned with persons is largely due to the tremendous impact Kant's theory of autonomy has had. In one sense or another, Kant is still in the background of the contemporary debate about personal autonomy.

Kant introduces the notion of autonomy as a key notion in his moral philosophy. For Kant, morality is a matter of a law that is universal and necessary. As rational beings, we are all bound by the moral law. At the same time, we are free. And Kant emphasizes that our freedom is a necessary condition for morality. This, however, creates a problem because at first glance there appears to be a tension between our freedom and our obligation to adhere to the moral law. It seems that insofar as we are subject to the moral law, we are not free, and insofar as we are free, we cannot be subject to the moral law. One of the most challenging problems of moral philosophy, then, is to give an account of moral obligation that gives credit to both requirements. According to Kant, all attempts to reconcile these fundamental aspects of morality have failed because they all turned the moral agent into a subject that has to abide by an external law, thereby violating the idea of freedom.

> "We need not now wonder, when we look back upon all the previous efforts that have been made to discover the principle of morality, why they have one and all been bound to fail. Their authors saw man as tied to laws by his duty, but it never occurred to them that he is subject only to *laws which are made by himself* and

[29] "The term's [autonomy] original meaning was political: a right assumed by states to administer their own affairs. It was not until the nineteenth century that 'autonomy' came (in English) to refer also to the conduct of individuals." Stephen Darwall (2006): 'The Value of Autonomy and Autonomy of the Will', in: *Ethics* (116), 263-284, 263. Darwall continues his analyzes by praising Kant for transferring the notion of autonomy to the personal level.

> yet are *universal*, and that he is bound only to act in accordance with a will which is his own but has for its natural purpose the function of making universal law."[30]

In this passage, Kant goes beyond a mere diagnosis of past mistakes and points us in the direction of a possible solution. The mistake has been to view the moral agent as a subject to a law that has its sources outside of the agent. Kant does not want to reject the idea that we are subject to the moral law. He deviates from previous attempts by putting forward the idea that the moral agent is not only subject to the law, but also its legislator at the same time. As free and rational agents, we are subject only to laws that we made ourselves. This is, in a nutshell, Kant's idea of autonomy.

Kant's major criticism of all accounts of morality that do not view the moral agent as a legislator of the moral law is that they make obedience to the moral law a contingent matter that depends on a further interest of the agent: "when they thought of man merely as subject to a law [...], the law had to carry with it some interest in order to attract or compel, because it did not spring as a law from *his own* will: in order to conform with the law his will had to be necessitated by *something else* to act in a certain way."[31] However, this violates the requirement that moral obligation is a necessary demand. For Kant, this problem can only be circumvented by a radical change in perspective. We need to let go of an understanding of morality that views moral obligation at bottom as an external demand and embrace a view of morality that empowers every moral agent to be a lawmaker. In Kant's words, we need to understand the moral agent not as being heteronomous, but as being autonomous.

> "This interest [to conform with the moral law] might be one's own or another's; but on such a view the imperative was bound to be always a conditioned one and could not possibly serve as a moral law. I will therefore call my principle the principle of **Autonomy** of the will in contrast with all others, which I consequently class under **Heteronomy**."[32]

[30] Immanuel Kant (1785/1965): *Groundwork of the Metaphysic of Morals*, trans. by H. J. Paton (New York: Harper Perennial), 100.
[31] Immanuel Kant (1785/1965), 100.
[32] Immanuel Kant (1785/1965), 100.

Kant introduces the notion of autonomy in order to ground morality. Morality depends on our autonomy, that is, on our capacity to bind ourselves by making universal laws. "Thus *morality* lies in the relation of actions to the autonomy of the will – that is, to a possible making of universal law by means of its maxims."[33] A necessary condition for autonomous lawmaking in this sense is that the will of the person is completely devoid of all empirical interests. Because she is autonomous, the moral agent transcends her entanglements in the realm of nature. "Autonomy of the will is the property the will has of being a law to itself (independently of every property belonging to the objects of volition)."[34] If it were otherwise, that is, if the law were in some way influenced by the particular interests of the agent, it would lose its claim for universal validity.[35]

An important consideration that is closely connected to what I have presented so far is that autonomy is, according to Kant, also the source of human dignity. Kant introduces his understanding of dignity in contrast to the notion of a price. "In the kingdom of ends everything has either a *price* or a *dignity*. If it has a price, something else can be put in its place as an *equivalent*; if it is exalted above all price and so admits of no equivalent, then it has dignity."[36] Everything that has a price can be traded, in principle, against something else. Something that has dignity, in contrast, can never be traded against something else – no matter how valuable the other thing is. According to Kant, all value finally depends on the existence of something that is an end in itself. "For nothing can have a value other than that determined for it by the law. But the law-making which determines all value must for this reason have a dignity – that is, an unconditioned and incomparable worth."[37] In other words, "that which constitutes the sole condition under which anything can be an end in itself has not merely a relative value – that is, a price – but has an intrinsic value – that is, *dignity*."[38] For Kant, the only end in itself is the autonomous person and, as such, she possesses dignity: "*Autonomy* is therefore the ground of the dignity of human nature

[33] Immanuel Kant (1785/1965), 107.
[34] Immanuel Kant (1785/1965), 108.
[35] This consideration leads Kant to the famous idea that the categorical imperative, which is the moral principle, is devoid of all matter and only concerned with form. "An absolutely good will, whose principle must be a categorical imperative, will therefore, being undetermined in respect of all objects, contain only the form of willing, and that as [sic] autonomy." Immanuel Kant (1785/1965), 112.
[36] Immanuel Kant (1785/1965), 102.
[37] Immanuel Kant (1785/1965), 103.
[38] Immanuel Kant (1785/1965), 102.

and of every rational nature."[39] Why is dignity rooted in autonomy? The reason is that rational agents, insofar as they make laws, give rise to value. That is, value has its ultimate source in autonomy. And this gives the bearer of autonomy dignity. "For it is not in so far as he is *subject* to the law that he has sublimity, but rather in so far as, in regard to this very same law, he is at the same time its *author* and is subordinated to it only on this ground."[40] Autonomous agents are authors of laws and thereby bring value into the world.

Kant's notion of autonomy is that of *moral* autonomy. Moral autonomy refers to the status of a person to be simultaneously legislator of and subject to the moral law. This is not the concept of autonomy that I am eventually interested in. If we think back to the examples I gave in the introduction, we see that the agents are autonomous in a different sense. What is remarkable about these agents is not that they possess moral autonomy, but that they possess *personal* autonomy.

Kant's major contribution to the concept of personal autonomy is that he took the idea of autonomy out of the political discourse and applied it to persons. The characteristics of personal autonomy become clearer in the next section when I engage in a discussion of Harry Frankfurt. An important difference between moral autonomy and personal autonomy is that moral autonomy in Kant's sense functions only by ignoring all individual characteristics of the agent, whereas personal autonomy is especially concerned with the interests, desires, and values that characterize the agent. Whereas Kant envisioned an autonomy that springs from pure reason and, as such, is devoid of all individual differences, the modern idea of personal autonomy emphasizes the particular characteristics of an agent's identity and her circumstances.

As we have just seen, Kant introduces the notion of autonomy in order to ground morality. This systematical interest does not fuel the modern debate any longer. The more individualistic concept of personal autonomy is not so much concerned with morality as it is with such things as well-being, authenticity, the ideal of a good life, and the value of diversity. Indeed, morality has not just lost its status as the raison d'être for a notion of autonomy. Autonomy is nowadays often conceived of

[39] Immanuel Kant (1785/1965), 103.
[40] Immanuel Kant (1785/1965), 107.

as being neutral with respect to morality. For Kant, this idea would be a contradiction in itself.

Why do I still view Kant as a milestone on the way towards the concept of autonomy that I am interested in? As I stated before, Kant was the first philosopher who systematically developed the notion of autonomy as a property of persons. And in Kant's own understanding, a person is autonomous insofar as she is not subject to external laws, but determines for herself what she ought to do. This contrast between external determination and self-determination forms the central idea of the concept of autonomy under investigation. An important difference is that Kant views the self as being devoid of empirical determination, whereas contemporary philosophers embrace the idea that a person develops as part of nature and that she forms her "practical identity," to use a phrase Christine M. Korsgaard has introduced,[41] in a context of constant influences. However, one can appreciate that the gist of the idea still remains similar: the autonomous person determines what she ought to do, and what she ought to do depends on who she is. Whereas Kant sees universal laws as the output of the workings of autonomy, the contemporary debate is more concerned with particular reasons for action. Few people still believe that individual autonomous choice necessarily leads to normative demands that can be universalized. This does not threaten morality because personal autonomy in this sense is not conceived as the core of a theory of morality.

1.3 Harry Frankfurt – Towards an Individualistic Understanding

Harry Frankfurt had a major impact on the contemporary debate about personal autonomy. He forcefully defended an individualistic understanding of autonomy, according to which a person is autonomous if and only if she identifies herself with her will. Identification is conceived of as a certain structure of the agent's desires. Frankfurt nullifies the role of morality and downplays the necessity of rationality for autonomous agency. He highlights, in contrast, what we might call volitional authenticity. In this picture, the autonomous standpoint of the agent is defined by her authentic self. With his hierarchical account of autonomy, Frankfurt developed a new center for the autonomy debate that virtually everyone who has worked on personal

[41] Christine M. Korsgaard (1996): *The Sources of Normativity* (Cambridge: Cambridge University Press).

autonomy in the last forty years or so has used as a starting point or target for developing their own thoughts. Frankfurt opened the room for our contemporary post-Kantian understanding of autonomy.

Before I discuss Frankfurt's hierarchical account in more detail, let me emphasize that my primary aim is to deepen our understanding of a specific concept of autonomy. Frankfurt added to this understanding. But, at the same time, it appears that he mixed it with other concepts such as free will, for example. The phenomenon to which he refers as free will in his early papers appears to be the same one that he discusses under the heading of autonomy later on. And in contrast to my understanding of free will, according to which free will is conceptually tied to alternative possibilities and the ability to choose otherwise, Frankfurt drops this assumption. My aim is to give a coherent interpretation of Frankfurt that fits the goal of explicating a particular concept of autonomy.

Frankfurt characterizes autonomy generally as follows: "An autonomous agent is, by definition, governed by himself alone. He acts entirely under his own control."[42] He explains what this means as follows: "A person acts autonomously only when his volitions derive from the essential character of his will."[43] In order to clarify the notion of an "essential character of will," Frankfurt uses his hierarchical account of desires and volitions, in particular his idea of internal and external desires. He characterizes autonomy by combining the idea of internal and external desires with the notions of being active and being passive:

> "Now insofar as a person's will is affected by considerations that are external to it, the person is being acted upon. To that extent, he is passive. The person is active, on the other hand, insofar as his will determines itself. The distinction between heteronomy and autonomy coincides, then, with the distinction between being passive and being active."[44]

[42] Harry G. Frankfurt (1999 b): 'Autonomy, Necessity, and Love', in: Harry G. Frankfurt (1999): *Necessity, Volition, and Love* (Cambridge: Cambridge University Press), 129-141, 132.
[43] Harry G. Frankfurt (1999 b), 132.
[44] Harry G. Frankfurt (1999 b), 133.

And this is the same phenomenon Frankfurt tries to explain earlier under the heading of free will. He addresses the same questions and develops his answers in the same framework. Indeed, he explicitly claims that "What really counts, so far as the issue of freedom goes, is not causal independence. It is autonomy. Autonomy is essentially a matter of whether we are active rather than passive in our motives and choices – whether, however we acquire them, they are the motives and choices that we really want and are therefore in no way alien to us."[45] In my interpretation, Frankfurt's work makes a lot of things visible that help us to understand autonomy more adequately. This is the background against which I will proceed in this section.

Frankfurt develops an account of autonomy that emphasizes what I want to call *self-directedness*. According to this picture, the autonomous agent is the self-directed agent, that is, the agent whose motives are in some special way expressive of herself or of her self. An agent fails to be self-directed, then, when her motives and actions somehow fail to be *her own* in an emphatic sense, which needs further clarification. Another way to put this distinction is by speaking of what an agent *really* wants in contrast to what she somehow might find herself inclined to do without really standing behind it or maybe even wanting to refrain from. When a compulsive gambler goes into the casino, even though she believes that this will lead to personal disaster and she tries as best as she can to refrain from gambling, she is, in a certain sense, not doing what she really wants to do. Many people are inclined to say that she goes gambling in spite of herself. But does it even make sense to say that the will that drives my action – a will that certainly is my will in some pretty straightforward sense – is in some sense not representing what I really want? For Frankfurt, the answer is an emphatic yes. His work is driven by the intuition that questions about autonomy are basically concerned with the distinction between self and non-self. Much of his work is inspired by the case of the unwilling addict and the question of how we can explain that the unwilling addict lacks autonomy. Let us then discuss this example in more detail.

Frankfurt describes the unwilling addict as follows: "[he] hates his addiction and always struggles desperately, although to no avail, against its thrust. He tries everything that he thinks might enable him to overcome his desires for the drug. But these desires are too powerful for him to withstand, and invariably, in the end, they

[45] Harry G. Frankfurt (2004): *The Reasons of Love* (Princeton: Princeton University Press), 20, FN 5.

conquer him. He is an unwilling addict, helplessly violated by his own desires."[46] According to Frankfurt, what is wrong with the unwilling addict is that he fails to endorse his motivating desire, or will, as Frankfurt also calls it. Since he is moved by a desire that he despises, he is not properly self-directed. And this is the reason why he is not autonomous in taking the drug. Frankfurt explains the intuitively compelling idea that the unwilling addict acts against his own will when he takes the drug by pointing out that the agent has a distorted psychological structure. In order to give a precise account of the psychological structure of the autonomous and the non-autonomous agent, Frankfurt introduces the notions of first-order desires and second-order volitions. Here is a more detailed description of unwilling addiction:

> "The unwilling addict has conflicting first-order desires: he wants to take the drug, and he also wants to refrain from taking it. In addition to these first-order desires, however, he has a volition of the second order. He is not neutral with regard to the conflict between his desire to take the drug and his desire to refrain from taking it. It is the latter desire, and not the former, that he wants to constitute his will; it is the latter desire, rather than the former, that he wants to be effective and to provide the purpose that he will seek to realize in what he actually does. […] The unwilling addict identifies himself […] through the formation of a second-order volition, with one rather than with the other of his conflicting first-order desires. He makes one of them truly his own and, in so doing, he withdraws himself from the other. It is in virtue of this identification and withdrawal, accomplished through the formation of a second-order volition, that the unwilling addict may meaningfully make the analytically puzzling statements that the force moving him to take the drug is a force other than his own, and that it is not of his own free will but rather against his will that this force moves him to take it."[47]

Autonomy, then, is a matter of having or lacking a certain kind of harmony in one's psychology. The unwilling addict lacks self-directedness and hence autonomy because he rejects the desire for taking the drug as his own, thereby exemplifying a

[46] Harry G. Frankfurt (1988 a): 'Freedom of the Will and the Concept of a Person', in: Harry G. Frankfurt (1988): *The Importance of What We Care About* (Cambridge: Cambridge University Press), 11-25, 17.
[47] Harry G. Frankfurt (1988 a), 17 f.

severe internal conflict. In Frankfurt's terminology, there is a clash between a "first-order desire" and a "second-order volition." Frankfurt defines these notions as follows: "'[F]irst-order desires' or 'desires of the first order,' […] are simply desires to do or not to do one thing or another."[48] First-order desires are directed towards action. Second-order desires, in contrast, are directed towards first-order desires. "Someone has a desire of the second order either when he wants simply to have a certain desire or when he wants a certain desire to be his will."[49] And Frankfurt further clarifies: "In situations of the latter kind, I shall call his second-order desire 'second-order volitions' or 'volitions of the second order.'"[50] Hence, second-order volitions are a special type of second-order desires: when the agent desires that a first-order desire should be her will, she has a second-order volition. To illustrate this terminology, let us consider an example.

The desire to take a drug is a first-order desire because taking a drug is an action. The desire to act on a desire to take a drug is a second-order volition because its content is that the first-order desire to take a drug should be one's will. If I want to be moved by a certain first-order desire, I have a second-order volition. Just wanting to have a certain first-order desire does not qualify as a second-order volition, although it is a second-order desire. That is, if I just want to have a desire for taking drugs, maybe because I want to know how it feels, but do not want to act on this desire, I have a second-order desire that is not a second-order volition. Since it is rarely the case that we desire to have a certain desire without also desiring to be moved by this very desire, we mostly need the distinction between first-order desires and second-order volitions for practical purposes.

Frankfurt focuses in his treatment of autonomous agency on the psychological make-up of the agent. He spells out the self vs. non-self distinction by distinguishing between specific kinds of volitional harmony and conflict. In contrast to Kant, who views the agent's self as determined by pure reason, and hence qualitatively identical for every agent, Frankfurt makes autonomy dependent on the authentic expression of the agent's individual characteristics. Each agent has a different volitional make-up that determines what is autonomous for her. Hence, with Frankfurt, the topic of what is characteristic for each individual agent becomes prevalent in the autonomy debate. I

[48] Harry G. Frankfurt (1988 a), 12.
[49] Harry G. Frankfurt (1988 a), 16.
[50] Harry G. Frankfurt (1988 a), 16.

address this topic of self-directedness more thoroughly in Chapter 5. There I also return to Frankfurt's hierarchical account of autonomy.

1.4 Conclusion

In this section I discussed three milestones in the history of the notion of autonomy. Autonomy started out as a political concept. For our discussion, it is of particular importance that autonomy was primarily an antagonistic notion, that is, it referred to situations of real or possible conflict and to a *polis*'s aptitude to deal with these conflicts. In other words, it referred to dispositions and abilities to prevail in conflicts. The notion of autonomy that I want to explore takes this aspect of the original understanding of political autonomy seriously. I will say more about the antagonistic understanding in Chapter 7. One of the key insights for the concept of personal autonomy under investigation is that autonomy is concerned with how an agent relates to other people, in particular, how she deals with social conflicts.

From Kant we take the focus on the personal with us. His account of moral autonomy is an essential background against which the contemporary discussion about personal autonomy unfolds. Although the concept I am interested in is in some respects very different from Kant's, Kant formulated a basic intuition about autonomy, namely that the autonomous agent follows her own agenda instead of letting herself be determined by external factors.

With Frankfurt we move even further, from the personal level to the individual characteristics of a person. Frankfurt actualizes Kant's idea in a way that opens up the route for an understanding of autonomy that focuses the authentic expression of a person's character.

This concludes my historical overview. I take with me a particular idea of autonomy, according to which autonomy is a property of persons and also that it is concerned on the one hand with how well a person prevails in conflicts and, on the other hand, with how authentically she expresses her individual character. In the next chapter, I expand on the issues in greater systematical detail.

2. The Concept of Autonomy

In the introduction, I presented examples of the kind of agency that forms the core of autonomous agency as I understand it. Antigone, Martin Luther, and Marie Curie all exemplify a particular concept of autonomy. These examples provide us with two hints about this particular idea of autonomy. First, autonomy in this sense is concerned with the expression of what a person regards as important. For Antigone, it is the proper burial of her brother; for Martin Luther, it is his religious conscience; and for Marie Curie, it is sciences. These paradigmatic instances of autonomy focus on agents who express what is of great importance to them. A second aspect that I regard as vital is that the examples are about situations of conflict. Antigone acted against the laws of the city and the commands of its ruler Creon. Martin Luther acted against the authority of the Catholic Church. And Marie Curie acted against the social expectation that made it difficult for women to become successful scientists. Taken together, a first tentative characterization of the concept I have in mind is as follows: autonomy is concerned with expressing what you regard as important *against opposition*.

The historical overview that I presented in the last chapter continued to look at these two aspects and sketched them further. I focused on three major developments: first, the emergence of autonomy as a political notion that referred to a *polis*'s assertiveness in dealing with other powers; second, Kant's groundbreaking work on transferring autonomy from the political to the personal level; and third, Frankfurt's move towards individuality and authenticity as the criteria for autonomy. The first step highlights the idea that autonomy deals with situations of conflict and, in particular, with an agent's ability to prevail in conflicts. The second step opens the route for a personal understanding of autonomy. And we can interpret the third step as exploring the idea that autonomy is concerned with an expression of what the agent regards as important.

In this chapter I want to follow up on these thoughts and sketch in more systematical detail the concept of autonomy that I explore and explicate in this study.[51] It is marked by the following aspects. First, autonomy is a property of person.

[51] The following characterization of the concept of autonomy relies in important aspects on Michael Pauen's conceptualization of autonomy. Especially Pauen's emphasize on the antagonistic character of autonomy, as I call it, that is, on the idea that autonomous agency is concerned with abilities to overcome obstacles is heavily reflected in my own understanding. Another important debt to Pauen's work consists in the dispositional reading of autonomy. Pauen argues for an understanding of

It can either be a local or a dispositional property. Second, it is a natural and contingent property in contrast to an essential and non-natural one. That is, human agents are not necessarily autonomous and autonomy supervenes on the natural characteristics of a person. Third, it is a gradual property and it can be practiced. Fourth, it is an antagonistic notion, that is, it is concerned with real or possible conflict and how the agent is disposed to deal with it. Fifth, it is concerned with how an agent forms her standpoint and how she expresses this standpoint in her life. It is conceptually related to authenticity. Sixth, it is morally neutral. Seventh, autonomy does not imply alternative possibilities. That is, it differs from free will. Let me discuss these aspects in turn.

2.1 A Property of Persons

The concept of autonomy that I explore in this discussion views autonomy as a property of persons. According to this understanding, saying, for example, that an action or decision is autonomous, is shorthand for expressing that a person is autonomous with regard to a particular action or decision. Understood as a property of persons, the notion of autonomy can be used in either of two ways: first, it can refer to a *dispositional property* of persons. With respect to this understanding, I will speak of *dispositional autonomy* or *global autonomy*. When we think of a person as being dispositionally autonomous, we think of her as a particular type of person, that is, a person who possesses certain abilities and dispositions. Second, the notion of autonomy can refer to a *local property* that a person has in a particular situation. In other words, we can refer to a person as being autonomous when she performs a particular action, makes a particular judgment, or behaves in a number of other ways *autonomously*. This is what I call *local autonomy*.

Dispositional autonomy is concerned with the more or less stable characteristics of a person and her patterns of choosing and acting. If we want to know whether a person is in general autonomous, we are concerned with dispositional autonomy. When we say that someone like Antigone or Marie Curie is an autonomous person, we employ the dispositional sense of autonomy. That is, we understand this person as someone who possesses certain dispositions and abilities that make up

---

autonomy according to which autonomy is a dispositional property of persons. I follow him in this. Compare Michael Pauen (2008): *Autonomy* (Manuscript).

autonomy. In my understanding, this dispositional reading of autonomy is the core one.

Local autonomy, by contrast, is concerned with datable instances of autonomous agency. We can also speak of the adverbial sense of autonomy because local autonomy is a specific way of doing something, and it is always possible to express this by saying that the person behaves autonomously. If we ask ourselves whether a person was autonomously voting for the Green Party, or whether someone chose autonomously to leave her family and start a new life, we invoke this local understanding of autonomy. In general, when we speak about autonomous actions and autonomous choices, we exploit the local sense of autonomy. As I pointed out, these expressions are shorthand for attributing autonomy to a person with respect to a particular action, choice, or other kind of behavior. If a person is acting under the guidance of someone or something external to her, she is not locally autonomous. Antigone's action of burying her brother was locally autonomous, as was Martin Luther's refusal to publicly renounce his opinions about the Catholic Church and the pope.

How local and dispositional autonomy are connected with each other is an open question. According to the understanding of autonomy that I explore, dispositional autonomy is the more basic phenomenon. By this I mean that we think of this kind of autonomy first and foremost in terms of dispositional autonomy. We might ask ourselves once in a while whether a certain decision, action, or the like has been autonomous. But the more typical question is how autonomous a person is globally conceived. The dispositionally autonomous person might lack local autonomy in a particular situation, but she typically displays a lot of autonomy in the way she leads her life.

2.2 A Natural Property

The concept of autonomy that I investigate conceives of autonomy as a natural property. To highlight this aspect, I also speak of *natural* autonomy.[52] According to

[52] The notion of natural autonomy is also used by Henrik Walter. I follow him in using it primarily to highlight an ontological position, namely the rejection of any non-natural elements in our ontological understanding of agency in general and autonomous agency in particular. Compare Henrik Walter (2001): *Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy* (Cambridge MA: MIT Press).

this understanding, autonomy is a property that supervenes on the physical world. The notion of natural autonomy refers to a set of dispositions and abilities, or their actualization, which constitute autonomy. The notion of natural autonomy allows us to conceive of human beings who are part of the fabric of nature as being autonomous without introducing any non-natural properties into our ontology. The contrast to natural autonomy is non-natural autonomy, which is made of non-natural properties. Kant's idea of moral autonomy is an example for an account of non-natural autonomy. A second and related aspect of natural autonomy is that it is a non-essential property. That is, it is not an essential property of persons to be autonomous. Whether a person is autonomous or not depends on her dispositional make up.

Accordingly, I investigate autonomous agency within a naturalistic framework. By this I mean, first and foremost, that I will restrict myself to a naturalistic ontology. This subscription to a naturalistic framework is rather the norm nowadays. However, non-natural concepts of autonomy exist. And given some of the central intuitions about autonomy, it might appear to be plausible to conceptualize autonomy in non-natural terms. I address this topic more thoroughly in Chapter 3. A related aspect is that an account of autonomy should be compatible with our best scientific account of the world. And it should be psychologically plausible. An account of autonomy should not make any assumptions about our psychological functioning that violate our knowledge about such things as psychological development, value-based decision-making, executive functions, and so forth. Natural autonomy is made up of specific dispositions and abilities which are concerned with how an agent develops her standpoint, forms her intentions, and executes them. These are also psychological subject matters. Hence, in order to avoid a misguided picture of autonomous agency, accounts of autonomy need to be psychologically plausible. It is, for example, psychologically implausible to assume that an agent's emotions have no impact whatsoever on her deliberations, practical judgments, and actions. Hence, a plausible explication of natural autonomy cannot see a contradiction between being emotionally engaged in one's practical endeavors and autonomy.

2.3 A Gradual Property

According to the concept of autonomy that I explore, persons can be more or less autonomous. Autonomy is, in other words, a *gradual* property of persons. Some

concepts of autonomy conceive of autonomy in binary terms. Think of Kant's concept of moral autonomy, for example. For Kant, every being endowed with reason is autonomous. This kind of autonomy does not come in degrees. One either has it or not. Another example for such a non-gradual concept of autonomy is the idea of autonomy as a right or source of rights. If we conceive of autonomy as a right, we usually don't want to operate with a gradual notion. After all, if autonomy is the source of the agent's status as a moral being, we should count every person as equally autonomous. Otherwise we would run into the problem that people who possess less dispositional autonomy have a lesser moral status. The understanding of autonomy as a the ultimate source of human rights has its roots in the Kantian theory of autonomy, according to which a person's dignity is grounded in her autonomy. We find reference to personal autonomy in this sense in discussions about human dignity and especially in many attempts to justify human dignity. A slightly different context in which autonomy is invoked in this right-sense is liberalism. Here, autonomy is often understood to be the right that paternalism violates.[53] The meaning of autonomy as a right is orthogonal to the notions of local and global autonomy. An agent might lack the latter ones but still retains the rights associated with autonomy. Even though it seems plausible to assume that there are some connections between these notions – a lot of philosophers would agree that autonomy as a right, in standard cases at least, presupposes the fiction of an agent who is dispositionally autonomous – autonomy as a right is especially important to ground basic forms of respect and moral agency. As such, it is less concerned with factual expertise or success. I am interested in autonomy as a natural property of persons. I won't have to say anything about the different concept of autonomy as a source of rights.

The gradual understanding of autonomy fits nicely with the dispositional reading of autonomy. It appears to be quite natural, for example, to think of agents as being more or less strongly disposed to follow other people's expectations. Dispositions such as these come in degrees. The same is true for the kind of abilities

[53] "One of liberalism's core commitments is to safeguarding individuals' autonomy." Joel Anderson/Axel Honneth (2005): 'Autonomy, Vulnurability, Recognition, and Justice', in: John Christman/Joel Anderson (eds.) (2005), 127- 149, 127. This "core commitment" goes back to Mill. "The classic argument for such an anti-paternalism constraint – from John Stuart Mill – is that for every one or two times somebody is stopped from doing something that he or she truly would have regretted, there will be dozens of interventions that simply represent other people's imposing their own conceptions of how best to live. Thus the net effect of a policy of paternalistic intervention will be an overall reduction in social welfare." Joseph Heath (2005): 'Liberal Autonomy and Consumers Sovereignty', in: John Christman/Joel Anderson (eds.) (2005), 204-225, 206 f.

that are constitutive of dispositional autonomy, for example, the ability to make up one's mind by impartially weighing reasons while being confronted with social pressure. Understanding autonomy in terms of dispositions and abilities gives us the conceptual space to conceive of it in a gradual fashion.

A related aspect is that autonomy can be fostered. In Western democratic societies, one goal of the educational system is to foster and enhance a child's autonomy.[54] The idea that autonomy can be developed, and that we can facilitate this process, is not limited to the education of children; it also plays a role for many adults who aspire to be more autonomous. We can learn to become more autonomous. We can strengthen our autonomy. Again, the dispositional reading of autonomy gives us a plausible explanation for this aspect of the concept of autonomy under discussion. When we understand natural autonomy in terms of certain dispositions and abilities, we have the resources to explicate how autonomy develops, how we can practice becoming more autonomous, and how we can foster autonomy in others.

2.4 An Antagonistic Notion

As has already become apparent in the historical overview, the concept of autonomy that I want to explore in this discussion is importantly linked to conflicts. An agent who goes her own way, that is, an agent who lives her life based on what is important to her, typically encounters a whole range of problems and difficulties in doing so. Autonomy is partly concerned with how well she handles these problematic and difficult situations, especially how well she deals with social pressure. The ability to overcome opposition and hindrances, more broadly conceived, is a hallmark of autonomous agency. This aspect of the concept of autonomy I am interested in is of particular importance because, although it fuels a lot of our intuitions about autonomy, it has received only scant attention in the contemporary debate.[55] I focus on this antagonistic dimension of autonomy in Chapters 6 and 7.

[54] Ishtiyaque Haji and Stefaan E. Cuypers explicate the idea that we aim at fostering our children's autonomy and discuss it in relation to an equal important value, namely well-being. They argue that fostering autonomy and well-being is importantly linked with each other. Ishtiyaque Haji/Stefaan E. Cuypers (2008): 'Authenticity-Sensitive Preferentism and Educating for Well-Being and Autonomy', in: *Journal of Philosophy of Education* (42), 85-106.

[55] As I mentioned above, Michael Pauen (2008) is a notable exception. Another philosopher who emphasizes that an aptitude to prevail in conflicts is constitutive of dispositional autonomy is Marina A. L. Oshana (1998): 'Personal Autonomy and Society', in: *Journal of Social Philosophy* (29), 81-102.

2.5 Expressing One's Own Authentic Standpoint

Autonomy in the sense under discussion is linked to the idea of authenticity.[56] The autonomous agent is able to shape her life in accordance with her own desires, beliefs, and values. And in order to explicate what this kind of ownership means, it is plausible to introduce a distinction between authentic and inauthentic desires, beliefs, and values. Authenticity is concerned with an agent truly being herself or being true to her innermost self.[57] The authentic person has a match between her feelings and her behavior. She does that which expresses who she is without being distorted by external influences.

Let me introduce a distinction between local and global authenticity. Local authenticity is concerned with particular instances of behavior. An agent is locally authentic if she is authentic in doing what she is doing in a particular instance, for example, smiling at her husband or expressing gratitude towards a friend. Global authenticity, in contrast, refers to the overall authenticity an agent expresses in her life. The globally authentic agent manages, by and large, to live her life authentically, although there might be episodes of local inauthenticity. A certain amount of global authenticity appears to be necessary for autonomy. But an agent can display autonomy also in inauthentic actions, or so it seems. Let us think of someone who is by nature submissive and always wants to please his wife. On a certain occasion, he suddenly feels reluctant to do as his wife tells him to do. He thinks for a moment and says to himself: "What the heck, I will just do something else." And he performs an action that is completely out of character for him. Now, I would think that the details of this example might be spelled out in a way that makes sense of the idea that someone acts inauthentically and yet autonomously.

This appears to be implied by her claim that an agent can only count as autonomous if she "can defend herself against (or be granted defense against) psychological or physical assault when it is necessary to do so." Marina A. L. Oshana (1998), 94. However, Oshana is not so much interested in the dispositions and abilities of the agent. Her focus lies more on the social context. Hence, she is not analyzing resolute agency in any detail.

[56] This position is defended, for example, by Insoo Hyun (2001): 'Authentic Values and Individual Autonomy', in: *The Journal of Value Inquiry* (35), 195-208;

[57] This is in accordance with the core meaning of the notion of authenticity as Charles Guignon analyzes it: "Built into this conception of autonomy is a distinction between what is really going on with me – the emotions, core beliefs, and bedrock desires that make me the person I am – and the outer avowals that make up my being in the public world." Charles Guignon (2008): 'Authenticity', in: *Philosophy Compass* (3/2), 277-290, 278. In the background of this understanding stands the idea that there is a meaningful way to distinguish between an agent's self and aspects of her that do not belong to her self. I discuss this idea in Chapter 5.

Although global authenticity is necessary for autonomy, it is by no means sufficient. Someone can be perfectly authentic when she acts conformist or submissive. It is not incoherent to imagine a person whose nature it is to do what others tell her to do. For this person, it would be inauthentic to withstand another person's demands or violate her expectations. I will come back to these questions in Chapter 6.

2.6 Not Tied to Norms – Morally Neutral

I have already made the distinction between personal autonomy and moral autonomy, viewing the latter as the non-natural property of persons who are, at the same time, the legislator and the subject of the moral law. Apart from this non-natural understanding of moral autonomy, we can also find tendencies to make a strong conceptual connection between morality and personal autonomy as the ability to govern oneself. One might think that self-governance requires the agent to be moral. Drawing somewhat loosely on Kant, one could assume that personal autonomy and morality are intimately connected in such a way that a person can only be autonomous insofar as she is moral. Christine Korsgaard's account comes close to this Kantian style picture of personal autonomy. I will discuss Korsgaard here more extensively because I will return later to her understanding of autonomy.

Korsgaard starts out with the question of where our obligations come from. What are "the sources of normativity"[58]? She argues that we have obligations because we are autonomous. Autonomy, in turn, refers to our ability to obligate ourselves. This idea of obligating oneself is familiar from Kant's concept of moral autonomy. Autonomy in this sense, one might think, does not refer to natural capabilities or contingent aspects of a person's identity, but to a non-natural property of persons. For Kant, and contemporary Kantians like Korsgaard, it is a non-natural fact that every person possesses moral autonomy. I do not want to deny that an understanding of autonomy as the "source of normativity" can be of great systematical value for some philosophical purposes. But this metaphysical usage does not help us to make sense of the idea that real people differ in their autonomy, that autonomy can be gained and lost, that we think of autonomy as an achievement, and so forth.

[58] Christine M. Korsgaard (1996).

Korsgaard is not only concerned with Kantian moral autonomy, but also discusses questions that pertain to the realm of personal autonomy. According to Korsgaard, an agent is autonomous if she governs herself on the basis of her judgments about what action is good. The foundation for these judgments is provided by the agent's conception of her own identity. "Autonomy is commanding yourself to do what you think would be a good idea to do, but that in turn depends on who you think you are."[59] As it stands, this appears to be compatible with immoral uses of autonomy. If morality plays no role in the "description under which you value yourself,"[60] then you won't take moral considerations into account when you deliberate about what to do. And this seems to be perfectly fine and poses no obstacle to your autonomy. But Korsgaard goes on by arguing that morality is a necessary part of one's practical identity: "moral identity is necessary."[61] Hence, at bottom, autonomy appears to require that the agent regard morality as important. Otherwise the agent would have no reason at all to do something. And this would exclude the possibility to "command oneself" because, as reflective beings, we need reasons for choosing a certain course of action. "The reflective mind cannot settle for perception and desire, not as such. It needs a *reason*. Otherwise, at least as long as it reflects, it cannot commit itself or go forward."[62] The thought, then, is this: without a reason, you cannot command yourself. And in order to have reasons, you need a practical conception of who you are. And your practical conception of yourself is necessarily built upon a self-understanding as a moral agent.

> "What makes morality special is that it springs from a form of identity which cannot be rejected unless we are prepared to reject practical normativity, or the existence of practical reasons, altogether […] Our practical identities depend for their normativity on the normativity of our human identity – on our own endorsement of our human need to be governed by such identities – and cannot withstand reflective scrutiny without it."[63]

[59] Christine M. Korsgaard (1996), 107.
[60] Christine M. Korsgaard (1996), 101.
[61] Christine M. Korsgaard (1996), 122.
[62] Christine M. Korsgaard (1996), 93.
[63] Christine M. Korsgaard (1996), 125.

Given this picture, you can only command yourself, i.e., be autonomous, if you understand yourself as a moral agent who is bound by moral norms. Hence, Korsgaard's notion of autonomy is an essentially moral one.[64]

In contrast to this, the understanding of autonomy that I want to explicate is compatible with non-moral exercises of autonomy. If the choices of a person lead her to actions that are not moral, she can still be autonomous. With regard to autonomy, the question is how she came to choose the immoral action or lifestyle. A mafia boss can be autonomous insofar as he makes his own choices and is able to defend them against opposition. His choices might be immoral, but they are still his choices.

Although I cut the connection between autonomy and morality, I am not opting for the opposite extreme, namely, that being moral contradicts an agent's autonomy. One might argue that fulfilling moral requirements or duties constrains an agent in a way that is incompatible with her autonomy.[65] But even though I reject the idea that there has to be a close marriage between autonomy and morality, I also reject the claim that morality necessarily obstructs an agent from being autonomous. Whether or not morality impedes an agent's autonomy depends on the agent. I do not think that this question can be answered in an *a priori* fashion. The understanding of autonomy that I want to shed light on and explicate more systematically is a subjective one, insofar as it views the agent's standpoint as the criterion for deciding whether or not she lives her life autonomously. If an agent takes moral standards seriously, if she cares about morality, then morality is an integral part of her standpoint. Thus, insofar as she managed to be autonomous, she will be moral. However, if an agent does not care about morality, she might be constrained in her autonomy by moral requirements.

[64] Let me mention, however, that Korsgaard seems to entertain doubts about her own claim that moral identity is necessary and that an agent cannot have any reason to act without committing herself to morality. In her "Reply" to a critical remark by G. A. Cohen, she concedes the possibility that an immoral Mafioso might have reasons to act and obligation to fulfill which have its source in his immoral practical identity. This opens the route for a conception of autonomy that is neutral with respect to morality. Christine M. Korsgaard (1996), 254 ff. I take up Korsgaard's notion of a practical identity in Chapter 5.

[65] A Nietzschean understanding of autonomy is an example for that.

2.7 Autonomy vs. Free Will

The concept of autonomy under discussion differs from the concept of free will. Sometimes these notions are used interchangeably or, at least, are treated as near equivalents.[66] However, as I understand it, autonomy differs from free will. Here is how I distinguish these two concepts. The will is free if the person could have chosen otherwise. That is, free will is concerned with our capacity to choose differently. Autonomy, in contrast, is concerned with governing oneself instead of being governed by someone or something else.

Let me expand on this a little more. As I understand it, the notion of free will is conceptually tied to the notion of alternative possibilities. This understanding of free will goes back at least to Aristotle who claims, "[…] when the origin of the action is in him, it is also up to him to do them or not to do them."[67] When we want to give an analysis of free will, we need to refer to the notion of alternative possibilities. The will of an agent is free if and only if she could have chosen to do something else. That is, freedom of the will requires that the agent can choose differently. Autonomy, in contrast, does not require the agent to have alternative possibilities. If an agent possesses the required dispositions, she is dispositionally autonomous. Every time she manifests these dispositions, she is locally autonomous. Neither local nor global autonomy require alternative possibilities because possessing and manifesting a disposition do not require alternative possibilities. For example, the disposition to strive for what you judge to be a worthy goal and a refusal to let yourself be intimidated by other people's demands are essential to the concept of autonomy under investigation. If an agent has this disposition and actualizes it, she is an autonomous agent who acts autonomously. There is no conceptual need to introduce a notion of alternative possibilities when analyzing autonomy and autonomous agency in this way. Moreover, that an agent has a free will does not exclude the possibility that she is disposed to do what others tell her to do. An agent can be free willed while at the same time lacking autonomy. It is not contradictory to imagine that someone who has the ability to choose differently acts conformist. The conformist is a paradigm

[66] Harry Frankfurt is an example of this as I pointed out in section 1.3.
[67] Aristotle (1985): *Nicomachean Ethics*, trans. by Terence Irwin (Indianapolis: Hackett Publishing), Book 3.

example of non-autonomy. However, even the conformist can enjoy freedom of the will.[68]

A second difference concerns the relations of autonomy and free will to responsibility. An agent can be responsible for her action even though she lacks local autonomy. After all, we do not want to let people off the hook just because they acted conformist.[69] Hence, local autonomy is not a necessary condition for responsibility. Maybe a certain level of dispositional autonomy is required for responsibility.[70] But free will certainly is different from dispositional autonomy. After all, an agent who is dispositionally autonomous might lack freedom of the will in a particular situation. And if free will is a necessary condition for responsibility, as many philosophers are inclined to think, it is meant as an actual property of the will that stands behind the action.

A third difference that I want to mention is that autonomy seems to be sensitive to substantial matters in a way that free will is not. It appears to be quite natural to view autonomy as being essentially concerned with choices and actions that have a high significance or importance for the agent. We might call them life choices and refer to the corresponding goals as "identity goals."[71] They include question like: should I marry or not? Should I have children or not? Do I want to work as a philosopher or as a social worker? What is more important to me: family or career? An agent's autonomy shows itself particularly in these existential matters. Apart from

[68] Let me clarify that conformism can be the consequence of an autonomous choice. That is, an agent might be locally autonomous, for example, in deciding to follow orders. We can think of someone who joins the army. The decision to join the army might be autonomous. However, blindly following orders is an example of non-autonomous agency. If the agent's commitment to a non-autonomous life-style is quite strong this even affects her dispositional autonomy because she is no longer disposed to think for herself, at least with respect to this domain of her life.

[69] It is plausible to assume that some kinds of conformism diminish or even nullify an agent's responsibility. Imagine cases in which the conformist behavior is the consequence of severe manipulation or indoctrination. However, there are a lot of examples for conformist behavior for which the agent is fully responsible. Think for example of a driver who drives too fast because all the others do it as well.

[70] Marina Oshana argues against the idea that autonomy of any kind is a necessary condition for responsibility. Her argument rests on two claims: first, an autonomous agent is not necessarily sensitive to moral reasons. Second, responsible agency requires an agent to be sensitive to moral reasons. Compare Marina A. L. Oshana (2002).

[71] Identity goals are goals which aim at being a certain person. Autonomy is especially concerned with the question of what kind of person I want to be as I discuss in Chapter 5. Strong identity goals facilitate resolute agency. I borrow the notion of identity goals from Gollwitzer. "Wicklund and Gollwitzer's (1982) research on symbolic self-completion more directly speaks to the issue at hand: people who were highly committed to achieving certain identity goals (e. g. becoming a successful musician) did not respond to failures, shortcomings, barriers or hindrances by retreat; rather, they stepped up their efforts to reach the intended goal." Peter M. Gollwitzer (1993): 'Goal Achievement: The Role of Intentions', in: *European Review of Social Psychology* (4), 141-185, 150.

rather bizarre circumstances, it would be very odd to point out that someone is autonomous in tying her shoes, brushing her teeth, or combing her hair. Even though we can imagine that these actions can also be performed non-autonomously, we usually do not judge a person's degree of autonomy by looking at them. To put it differently: if we compare two agents, and one of them is autonomous in choosing her occupation and in engaging in her family life, though she lacks autonomy with regard to brushing her teeth and tying her shoes, and the other agent is autonomous in brushing her teeth and tying her shoes, but lacks autonomy in her choices about what to do for a living and whom to live with, we would certainly view the first person as being much more autonomous. Free will, however, is a different matter. If my will is free, I am as free willed in tying my shoes as I am in deciding whom I want to marry. I am not showing more freedom of the will in the latter case, although the decision is much more important. Of course, freedom of the will might matter for us especially when we think about very important matters. But in contrast to autonomy, there is no sensitivity to substantial matters built into the notion.

A last distinction between free will and autonomy, as I understand it, is that the latter is, whereas the former is not, a matter of degree. An agent can be more or less autonomous. On the global level, this is relatively easy to see because agents tend to diverge in their autonomy in different contexts or at different times. But local autonomy is also a matter of degree. External influences can be more or less penetrating. Someone might reject a certain demand, thereby expressing some autonomy, while she still acts within the social constraints that she has uncritically internalized. An agent's will, on the other hand, is either free or not. Either an agent can choose differently or not.

Some people might object to this thought by pointing out that we can think of free will also as a gradual property. First, we might think that an agent enjoys the more freedom of the will the more options she has to choose from. This would be a quantitative criterion. A second, qualitative criterion is concerned with how significant the agent's options are. The choice between different flavors of a milk-shake and the choice between different occupations differ in their significance – the latter one being the more significant one. According to the significance criterion, then, an agent enjoys more freedom of the will if the choice of her occupation is open to her compared to the agent who can only choose between different flavors of her milk-

shake. A third possible criterion refers to the epistemic position of the agent. According to the epistemic criterion an agent is the more free-willed the more she knows about the relevant consequences that are attached to her options.

The quantitative criterion seems to me a rather implausible one if it is not accompanied by a standard of significance. That is, from an intuitive point of view more options appear only to enhance one's degree of freedom of the will, if at all, when they have real significance for the agent. Why should my degree of freedom of the will be higher when I can chose to pick up one of 1000000 grains of sand compared to the agent who can chose between 10000 grains of sand? Does it follow that we should add a significance criterion to our notion of free will? A little above I argued that autonomy is, whereas free will is not essentially concerned with the significance of an agent's choices and actions. If we introduce a significance criterion into our concept of free will we would confound the concepts of autonomy and free will. As I initially said, philosophers did this regularly. However, for reasons of conceptual clarity we should distinguish between free will and autonomy. Hence, I understand free will, in contrast to autonomy, as being unaffected by matters of significance. Finally, regarding the epistemic criterion I want to emphasize that the idea of being able to choose otherwise is not conceptually tied to an agent's epistemic position. Of course, we could still add this criterion to our conception of free will. If you are inclined to do this, then free will also becomes a gradual concept. We should note, however, that we would still operate with two different standards of measuring an agent's degree of freedom of the will and her degree of autonomy. That autonomy comes in different degrees is primarily a matter of an agent's dispositions and abilities. The epistemic criterion, in contrast, refers to an agent's knowledge. Hence, we are not in danger of confounding the two notions. An agent's will is free if and only if she is able to choose otherwise. An agent is autonomous if she is able to express what is important to her against opposition.

2.8 Some More Remarks

In the preceding sections, I have outlined the concept of autonomy that I am interested in. According to this concept, autonomy is a natural property of persons that comes in different degrees and is concerned with authentically expressing one's own standpoint and prevailing in conflicts. It is neither tied to norms nor does it presuppose

alternative possibilities. Before I investigate these issues in greater depth in the next chapters, let me clarify some further issues. Sometimes autonomy is equated with autarky. Autarky is complete independence from other people. The autarkical person relies only on herself. In contrast, autonomy as I understand it is compatible with reliance on others. However, the belief that autonomy and dependence contradict each other is rather pervasive. Susan Wolf, for example, understands autonomy in terms of absolute independence from any influences whatsoever: "there is a requirement that the agent's control be ultimate – her will must be determined by her self, and her self must not, in turn, be determined by anything external to itself. This last condition I shall call, after Kant, the requirement of autonomy."[72] Unsurprisingly enough, she concludes that in the light of this understanding, autonomy cannot be an ideal for us since we are all dependent creatures. "The idea of an autonomous agent appears to be the idea of a prime mover unmoved whose self can endlessly account for itself and for the behavior that it intentionally exhibits or allows. But this idea seems incoherent or, at any rate, logically impossible."[73] Of course, Wolf is right in claiming that the completely self-reliant agent fundamentally differs from us. However, she obviously uses another concept of autonomy when she explicates it in this way. The autonomous person can be influenced by other people, according to the concept of autonomy under discussion.

Advice is a good example of an influence that does not undermine autonomy. The agent remains in control regarding how she uses the advice. Her critical engagement in the question she fights with is not subdued by getting advice. Of course, it is possible that she just follows the advice without giving it a second thought. In this case, she might damage her autonomy. But at the moment we are concerned with the possibility of being influenced without losing autonomy. And getting advice quite often exemplifies this possibility.

Practical projects often need the support of other people. A violin player who wants to play a violin concerto requires a whole orchestra. The success of her project depends on dozens of other people. This in itself, however, does not deprive her of autonomy. It is perfectly possible that she guides herself in accordance with what she finds valuable. And this is what matters. If it were otherwise, our autonomy would

[72] Susan Wolf (1990), 10.
[73] Susan Wolf (1990), 14.

indeed be in jeopardy because we all depend in so many ways on the labor, support, and participation of others that we would be rendered non-autonomous in virtually all of our endeavors. Another sort of external influence that is sometimes construed as an obstacle to autonomy is causal influence. Again, if causal influence were to nullify our autonomy, we certainly would lack autonomy because we are natural beings and, as such, part of the causal nexus. Such an understanding of autonomy has its source in a different concept. I investigate natural autonomy, and natural autonomy is not necessarily undermined by causal influence.

The question with respect to autonomy, then, is not whether an agent is externally influenced by or dependent on others. The question is how she is influenced and in which ways she depends on others. Manipulation and coercion, for example, are autonomy-undermining influences. Advice and help, on the other hand, are not. Understanding autonomy in this way allows us to retain our self-understanding as autonomous agents. If autonomy were identical to, or dependent on autarky, no human being would ever be autonomous.

Another qualification concerns the relation between rationality and autonomy. It was Kant who emphasized the essential role of rationality for autonomy. According to Kant, an agent is autonomous if and only if she acts rationally. The contemporary debate still conceives of rationality as an important, maybe even essential aspect of autonomy. Korsgaard, for example, argues that autonomy becomes an issue for us only because we are rational agents who do not act simply on impulse, but are guided by reasons. And we are autonomous only insofar as we act for reasons that express our practical identity.[74] However, the emphasis on rationality has also sparked some skeptical comments. Keith Lehrer, for example, claims that being under the command of rationality might pose a threat to autonomy:

> "Reason has co-opted our conception of autonomy. My purpose is to set autonomy free. Here is the problem: some philosophers, most notablyKant, have said that governing your life by reason or by being responsive to reason is the source of autonomy. But there is a paradox concealed in these plausible claims.

[74] Christine M. Korsgaard (1996).

> […] a person can be enslaved to reason and lack autonomy because of this kind of bondage."[75]

Before I comment on this issue, I need to make some terminological clarifications. First, I entertain a fairly broad notion of rationality, according to which someone qualifies as rational if she is sensitive to reasons. The contrast to rationality in this sense is arationality, that is, a form of agency for which reasons play no role. Second, I want to distinguish between theoretical and practical rationality. Theoretical rationality is concerned with the formation of beliefs. That is, theoretical rationality is a doxastic matter. Practical rationality, on the other hand, is concerned with an agent's practical endeavors, that is, the formation of her intentions and her actions. Third, we can distinguish between local and global rationality. An agent is locally rational if she fulfills the requirements of rationality in a particular situation with respect to the formation of a belief (if it is theoretical rationality) or the formation and execution of an intention (if it is practical rationality). Global rationality, on the other hand, refers to the agent's dispositions and abilities that allow her to rationally form beliefs (theoretical rationality) or to rationally choose and act (practical rationality). A globally-rational person might be locally irrational in a particular situation. And a locally-rational person might be globally irrational.

The concept of autonomy that I am interested in makes global theoretical as well as global practical autonomy a necessary condition of autonomy. That is, no agent who completely lacks the dispositions and abilities that allow her to rationally form beliefs and to rationally choose and act can count as autonomous. Local theoretical irrationality, however, might be compatible with local autonomy.[76] The same is true for local practical irrationality. An agent is locally practically irrational if and only she acts against her better judgment or if her judgment is based on faulty

[75] Keith Lehrer (2003): 'Reason and Autonomy', in: Ellen Frankel Paul/Fred D. Miller, J./Jeffrey Paul (eds.) (2003): *Autonomy* (Cambridge: Cambridge University Press), 177-198, 177. In a similar vein, Susan Wolf says: "the ability to act in accordance with Reason may seem to free us from one threat to autonomy only at the cost of making us susceptible to another. […] if we cannot help choosing the most rational alternative, we are not autonomous agents." Susan Wolf (1990), 51 ff.

[76] It seems *prima facie* plausible that I can be autonomously performing a certain action even if the reasoning that leads me to this action was not perfectly theoretically rational. Let us imagine that I oversaw a logical implication of a belief of mine and that I would have decided to act differently if I had seen this implication.

practical reasoning such as wishful thinking. Where does this leave us with respect to skeptical worries Lehrer formulates?

I think that it is a fundamental mistake to view rationality as a threat to autonomy. First, I reject one of the premises that apparently grounds such a claim, namely that autonomy is identical to, or requires some form of absolute independence or autarky. I already pointed out why I reject the equation of autonomy and autarky. In contrast to such an approach, I want to understand autonomy as a property of dependent creatures like us. I am interested in what I have called natural autonomy. The question with regard to an agent's autonomy, then, is not whether she is constrained, but how she is constrained. Some constraints undermine autonomy, some do not.

Second, I claim that rationality is a necessary condition for autonomy. Doing what one thinks one has good reason to do, and doing it because one thinks so, does not threaten or thwart the agent's autonomy. Being rational in this way is a necessary condition for autonomous agency as I understand it because an agent acts autonomously only if she acts for a reason. Her reasons define her autonomous standpoint. I argue for this understanding in Chapter 5.

2.9 Conclusion

In this chapter, I have sketched in more systematical detail the concept of autonomy that I explore throughout this analysis. According to this understanding, autonomy is a natural and gradual property of persons, it is not tied to norms, and it differs from free will. It deals with an agent's authentic expression of her standpoint and her aptitude in prevailing in conflicts or difficult situations. My aim now is to explicate these aspects of autonomy more thoroughly. I will start in Chapter 3 with explicating the idea that autonomy is a natural property that needs to be accounted for in a naturalistic framework.

3. Natural Autonomy

In the last chapter, I sketched the concept of autonomy that I want to explore and explicate more thoroughly in this study. A central aspect of this concept is that it conceives of autonomy as a *natural* property. Hence I also speak of natural autonomy. When we use this notion, we highlight that we are concerned with natural abilities and dispositions of agents who are part of the physical world. Natural autonomy, in other words, is attributed to us as biological creatures that completely belong to the physical realm. This obviously contrasts with a Kantian notion of moral autonomy, according to which autonomy has its source in our pure reason and which, in turn, is not part of our empirical make up, but rather a non-natural property of persons. We also came across other concepts of autonomy that contradict the idea of natural autonomy. Think of Susan Wolf's understanding of autonomy again, according to which it is a "requirement of autonomy" that the agent is in control of her will and that this control is "ultimate," that is, "her will must be determined by her self, and her self must not, in turn, be determined by anything external to itself."[77] As natural beings, we are shaped in manifold ways by things external to us. Natural autonomy is autonomy within nature.

In this chapter, I explicate the ontological requirements that the concept of natural autonomy has to fulfill. Natural autonomy is an extension of our natural ability to act. Hence it makes sense to start an investigation of natural autonomy on the basis of a naturalistic account of agency and action. As I have stated, this chapter addresses the issue of naturalism. What do I mean by referring to an account of agency and action as a *naturalistic* one? As I understand it, this is primarily an ontological matter. In particular, an account is a naturalistic one, in my sense, if and only if it stays within an event-causal framework. This is the topic of this chapter. In Chapter 4, I will build on this foundation and investigate how we can conceptualize agential control and account for agency and action within a naturalistic framework.

3.1 Naturalistic Framework

There is no shared agreement about what makes a theory naturalistic. David Papineau observes that "[t]he term is a familiar one nowadays, but there is little consensus on its meaning. For some philosophers, the defining characteristic of naturalism is the

[77] Susan Wolf (1990), 10.

affirmation of continuity between philosophy and empirical science. For others the rejection of dualism is the crucial requirement. Yet others view an externalist approach to epistemology as the essence of naturalism."[78] I understand naturalism basically as an ontological matter. Hence, for me, the rejection of dualism is a central aspect.

Additionally, the scientific investigation of agency, in general, and autonomous agency, in particular, is not a project completely distinct from the subject matter of a philosophical account. Moreover, for some of the issues in the philosophy of action, empirical investigations are directly relevant.[79] I view it as a requirement for naturalistic accounts of agency that they are compatible with our best scientific theories about the world in general and agency in particular. This constraint is especially important for demanding forms of agency such as autonomous agency that attribute a complex set of dispositions and abilities to the agent. For example, the autonomous agent possesses, among other things, particular rational skills and a particular form of independence from external influences. Now, we are all heavily influenced by other people and the situation we are in. Such disciplines as social psychology and cognitive neuroscience provide a plethora of data that show what kinds of influences have an impact on our desires, choices, and actions.[80] If an account of autonomy does not pay attention to these findings, it is in serious danger of being completely irrelevant because it entertains mistaken beliefs about the workings of human agency. We have, for example, increasing evidence on the fundamental role

[78] David Papineau (1993): *Philosophical Naturalism* (Oxford: Blackwell Publishers), 1.

[79] Here are just two examples. First, there is increasing empirical evidence that a lot of our behavior is automatically caused and not under immediate conscious control. (Compare for example the comprehensive overview in Daniel M. Wegner (2002): *The Illusion of Conscious Will* (Cambridge, MA: Harvard University Press).) These findings are directly relevant for a philosophical account of how we guide and control our actions. (Compare for example Alfred R. Mele (2009): *Effective Intentions. The Power of Conscious Will* (Oxford: Oxford University Press).) A second example concerns the phenomenology of action. Based on empirical research on schizophrenic patients, philosophers of action have started to distinguish between different kinds of phenomenological experiences on the side of the agent. In particular, the distinction between a sense of authorship and a sense of ownership has become standard fare. And the phenomenological discussion also has an impact on theories about the underlying mechanism of agency and action. Compare for example Chris D. Firth/Sarah-Jayne Blakemore/Daniel M. Wolpert (2000): 'Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action', in: *Brain Research Reviews* (31), 357-363; Elisabeth Pacherie (2008): 'The Phenomenology of Action: A conceptual framework', in: *Cognition* (107), 179-217; Tim Bayne: 'The Sense of Agency', in: Fiona Macpherson (ed.) (2011): *The Senses. Classic and Contemporary Philosophical Perspectives* (Oxford: Oxford University Press), 355-374.

[80] The most famous example from social psychology for the influence of social pressure certainly is the Solomon Asch experiment (Solomon Asch (1956): 'Studies of Independence and Conformity. A Minority of one against an unanimous majority', in: *Psychological Monographs* (70), 1-70).

of emotions in our deliberations and choices.[81] It appears to be a contingent fact about human agents that they are severely impaired in their rational capacities if they are unable to feel emotions. Given that this is true, the dispositions that constitute autonomy need to include emotional capacities. My claims about autonomous agency are partly derived from empirical considerations and, for that matter, defeasible by empirical findings.[82]

As I have made clear, I am mainly concerned with naturalism as an ontological approach. That is, when I speak about a naturalistic account of agency and action, I have an account of agency and action in mind that stays within the limits of a naturalistic ontology. Let me flesh out in more detail the ontological requirement for an account of agency that aspires to be naturalistic. As I have highlighted, following Papineau, the rejection of dualism is a requirement. Any attempt to account for agency in terms of a dualistic ontology in which agency is located in some non-natural substance fails to be naturalistic. The most influential example of a dualistic account of agency is Descartes' philosophy. Descartes assumes that there are two different substances, *res cogitans* and *res extensa*.[83] The first constitutes the physical world. The second is an additional, non-natural substance out of which our souls are made. This kind of approach falls outside of the scope of a naturalistic framework.[84] For a naturalistic understanding of agency, recourse to non-naturalistic substances in which agency resides is blocked. Any component in our ontology that transcends the natural world is by definition non-naturalistic.

The rejection of dualism is hardly a surprise for any contemporary philosophical endeavor. This raises the question of why we should consider a non-natural understanding of autonomy in the first place. What initial plausibility has such an understanding? The strongest reason, to my mind, is the idea that the autonomous

---

[81] A hallmark in this discussion about the essential role emotions play in deliberation and decision making is Antonio R. Damasio (1994): *Descartes' Error. Emotion, Reason, and the Human Brain* (New York: G. P. Putnam's Sons). Compare from a philosophical point of view: Paul Thagard (2001): 'How to Make Decisions: Coherence, Emotion, and Practical Inference', in: Elijah Millgram (ed.) (2001): *Varieties of Practical Reasoning* (Cambridge MA: MIT Press), 355-371.

[82] The whole understanding of resolute agency that I develop in Chapter 7 is strongly influenced by empirical research on intentional action and the phenomenon of willpower.

[83] René Descartes (1641/1986): *Meditations on First Philosophy*, trans. by John Cottingham (Cambridge: Cambridge University Press).

[84] Gilbert Ryle's polemical attack against the "ghost in the machine" certainly is one of the most important criticisms of Descartes' substantial dualism in the 20th century: Gilbert Ryle (1949): *The Concept of Mind* (London: Hutchinson). Compare also Colin McGinn (1982): *The Character of Mind* (Oxford: Oxford University Press).

agent determines her own fate, instead of being determined by external influences. If I am supposed to be the author of my life, it appears to be problematic if my actions are caused by events beyond my control. I am only autonomous, according to this intuition, if I control my actions. Hence, any external causal determination threatens my autonomy. But as part of nature, I appear to be externally determined. Hence, my autonomy must have its source in some non-natural part of me, for example, in a non-natural substance like Descartes' *res cogitans* or in Kantian pure reason. This kind of reasoning illustrates the attractiveness of a non-natural account of autonomy. I will develop this thought in more detail in the following section about the challenge from missing control.

The naturalistic requirement blocks the introduction of any non-natural substances or entities, such as immortal souls or the like. In addition to that, it also excludes another prominent way of accounting for agency and action, namely, by assuming a kind of causality that is distinct from event-causation and specific to action. Some philosophers have argued that agency as such requires the ability to initiate causal chains whereby this initiation is not itself naturally caused, but originates in the agent. The metaphor of the unmoved mover illustrates this intuition. Philosophers introduced the notion of *agent causation* for this kind of causality. C. D. Broad explicates agent-causation as the idea that action is "literally determined *by the agent or self*, considered as a substance or continuant, and not by a total cause which contains as factors *events in* and *dispositions of* the agent. If this could be maintained, our puttings-forth would be completely determined, but their causes would neither be events nor contain events as cause-factors."[85] Roderick Chisholm's notion of "immanent causation" is similar.[86] The rationale behind the idea of agent causation is the intuition that we wouldn't have control over our actions, and wouldn't be responsible for them, in turn, if our actions were caused by a chain of events that transcends the agent. The notion of agent causation, then, attempts to give an answer to the question how agents have control over their actions. And this also makes it an attractive candidate for explicating the special control autonomous agents exert in shaping their lives. However, when we presuppose agent-causation we violate the naturalistic requirement.

[85] C. D. Broad (1952): 'Determinism, Indeterminism, and Libertarianism', in: C. D. Broad (1952): *Ethics and the History of Philosophy* (London: Routledge and Kegan Paul), 195-217, 214 f.
[86] Roderick M. Chisholm (1966): 'Freedom and Action', in: Keith Lehrer (ed.) (1966): *Freedom and Determinism* (New York: Random House), 11-44.

Agent causation is a kind of causation that differs from that form of causation that governs the realm of the physical. As John Bishop observes, "[n]aturalism does not essentially employ the concept of a causal relation whose first member is in the category of person or agent (or even, for that matter, in the broader category of continuant or 'substance'). All natural causal relations have first members in the category of event or state of affairs."[87] I assume that event causation is the default notion regarding causal connections in the natural world. Event causation is a relation between events whereby one event causes the other. There are different ways to spell out this relation. What they have in common is the assumption that events are always caused by other events and that no event originates without a cause. According to an event causal picture, events "occur in accordance with deterministic or probabilistic laws, given antecedent (and perhaps also, concurrent) events and states of affairs."[88] Hence, an agent's causing her action is an event that itself is caused by other events. Agent causation, as a form of causality in which an agent causes a new causal chain without being caused to do so, does not fit into this picture. Of course, this is unproblematic for proponents of agent causation because they think that agency cannot be accounted for in an event-causal framework. The challenge, then, for a naturalistic account of agency, consists in explaining agency without reference to agent causation.

A related notion, which is not available to naturalistic accounts of agency, is that of a *volition* understood as an irreducible mental act. Berent Enç gives a good general description why one might be inclined to introduce volitions into one's action theory.

> "[T]he predominant way of distinguishing voluntary action from non-voluntary ('mere') behavior consists in identifying some special class of causes located within the agent, and stipulating that voluntary action is marked by the effective role played by such causes. This way of approaching the problem forces an immediate choice between two options on the theorist. One option is to designate some mental *action* as the cause. The second option, in a more reductionistic

[87] John Bishop (1989): *Natural Agency. An Essay on the Causal Theory of Action* (Cambridge: Cambridge University Press), 40.
[88] John Bishop (1989), 39.

> spirit, is to analyse actions into behavior caused by some special class of *mental events* that are not themselves actions."[89]

And with respect to the first option, Enç points out that "[t]he favorite mental act in this context is an act of will, or a volition – acts that are thought to be essential to agency in that agents always initiate their actions 'through' them."[90]

Volitions violate the naturalistic requirement because they are not accountable in event causational terms. Of course, we might define volitions as mental actions that have an event causal history. However, in this case we wouldn't have made any progress towards a naturalistic account of agency since we still need an account of the mental causes of actions, although now we direct our attention to mental actions. Irreducible mental acts violate the ontological requirement because they do not explain agency and action within a naturalistic ontology, but instead introduce a peculiar kind of action into our basic ontology.

A theory of agency and action is naturalistic in my sense, then, when it is committed to a naturalistic ontology, that is, when it is explicated purely in event-causal terms. This implies the rejection of dualism, agent causation, and volitions as irreducible mental actions. Therein lies a challenge, which I call, following John Bishop, the problem of natural agency: "the problem of natural agency is an ontological problem – a problem about whether the existence of actions can be admitted within a natural scientific ontology."[91] If we cannot solve this problem, we get a fragmented view of ourselves that incorporates contradictory elements. This feels deeply unsatisfactory. The aim, then, is to show that we are autonomous agents within nature – a position that Bishop calls "reconciliatory naturalism."[92] Reconciliatory naturalism assumes that "the presuppositions of our ethical and natural scientific perspectives are, in fact, mutually consistent."[93]

I shall simply assume that realism about mental states is true and that there is, in principle, a way to explicate how mental states fit into the physical world, in

[89] Berent Enç (2003): *How We Act. Causes, Reasons, and Intentions* (Oxford: Oxford University Press), 6.
[90] Berent Enç (2003), 6.
[91] John Bishop (1989), 40.
[92] John Bishop (1989), 5.
[93] John Bishop (1989), 5.

particular, how they can have a causal impact. Hence, the worry I address is not a general skepticism about mental causation, but skepticism about the idea that causation by mental states somehow adds up to performing an action. This kind of skepticism is based on two considerations that both provide a challenge for any attempts to develop a naturalistic account of agency and action. The first challenge is the challenge of missing control, and the second is that of the missing agent. I present them both in turn.

3.2 A First Challenge: The Challenge of Missing Control

It is a common assumption that having control over one's behavior is a necessary condition for action.[94] An agent who acts exercises control in a way that is absent in mere happenings in which the agent is involved. We can explicate this intuition by saying that the agent is the active subject of an action and not the passive object of some unfolding events. The challenge of missing control says that control vanishes in an event-causal description. In other words, a fundamental challenge for causal theories of action is to explicate how we can make sense of the idea that an agent is the source of her actions, i.e., that she has control over her actions in purely event causal terms. Peter van Inwagen's consequence argument is one way to phrase this challenge. He says, at bottom, that behavior that is the deterministic consequence of things beyond the agent's control is itself not under the control of the agent.[95] Here is a more extensive way to formulate the consequence argument:

> "If an event that constitutes an agent's action is causally determined, it occurs as a deductive consequence of relevant laws of nature and prior states and events. But neither of these determining factors is under the agent's control. Still, to count as the agent's own action, what happens must be an exercise of *the agent's* control. Therefore, the very same event must both occur under the agent's control and also be determined by factors that are beyond the agent's control. But this is impossible. Hence, no action can also be causally determined."[96]

[94] In the next chapter I argue for an even stronger claim, namely the claim that actions just are exercises of control by the agent.
[95] Peter van Inwagen (1983): *An Essay on Free Will* (Oxford: Clarendon Press), Chapter 3.
[96] John Bishop (1989), 26. Of course, Bishop does not defend the consequence argument, but reconstructs it in order to explicate the challenge it poses.

The underlying thought is that circumstances that are not under our control determine what we do. Hence, what we do cannot be under our control.

John Martin Fisher and Mark Ravizza discuss this challenge quite extensively in their attempt to develop the notion of agential control that is essential to their understanding of responsibility. According to them, one way to explicate this challenge to control exploits "the Principle of the Transfer of Powerlessness": "the principle says that if *p* obtains and a person *S* cannot so act that *p* would be false, and *S* cannot so act that it would be false that if *p* then *q*, then *q* obtains and *S* cannot so act that *q* would be false."[97] This is indeed a quite plausible principle. In order to use it for making the case against the possibility of agential control, it is combined with two other principles. The first one, which Fisher and Ravizza dub "the Principle of the Fixity of the Past," states, "that if a person's performing a certain action would require some actual fact about the past not to have been a fact, then the person cannot perform the act."[98] The second one is called the "Principle of the Fixity of the Laws." It claims that "if performing a certain action would require that some actual natural law *not* be a law, then the person *cannot* perform the act."[99] Assuming that these three principles hold, we can conclude that we don't have control over what we do because we can change neither the past nor the laws of nature. At least, we don't have a certain kind of control that is concerned with choosing between different alternative courses of action. Fisher and Ravizza describe it in the following way: "When we take one path rather than another in a situation in which the other path is genuinely available to us, we say that we have a certain kind of *control* over our behavior. In this kind of circumstances, a person has the sort of control that involves alternative possibilities."[100] Now, if determinism is true, and if we acknowledge the three principles that I just mentioned, then we lack this kind of "alternative-possibilities control."[101] This is a serious challenge for a naturalistic account of agency and action. It forces us to develop a notion of control that is strong enough to be used as the

[97] John Martin Fisher/Mark Ravizza (1998): *Responsibility and Control. A Theory of Moral Responsibility* (Cambridge: Cambridge University Press), 18.
[98] John Martin Fisher/Mark Ravizza (1998), 19.
[99] John Martin Fisher/Mark Ravizza (1998), 19 f.
[100] John Martin Fisher/Mark Ravizza (1998), 20.
[101] John Martin Fisher/Mark Ravizza (1998), 31.

conceptual foundation of an account of action while still remaining within a naturalistic ontology. I pursue this task in Chapter 4.

3.3 A Second Challenge: The Missing Agent

The rejection of dualism, agent-causation, and volitions poses a second and related challenge for naturalistic accounts of agency and action since it is intuitively attractive to assume that the control we have over our actions does not reduce to the occurrence of some events in us. Naturalistic accounts of agency and action share the basic assumption that an action is behavior that is caused, in the right way, by the agent's mental states. At first glance, however, it might seem that, just as I remain passive when my immune system reacts to a virus, I am also passive when some mental states of mine cause some behavior. We can call this the challenge of the missing agent. The challenge arises because naturalistic accounts of agency and action need to abandon reference to the agent and her doings in their explanation of agency and action. Otherwise they would be circular. But when they leave out reference to the agent and what she does in accounting for agency and action, there might be no way to draw the agent back into the picture.

David Velleman forcefully gives voice to the intuition that approaches that understand actions as behavior with a specific mental etiology are abandoning the agent. He sketches what he calls "the standard story of human action,"[102] according to which, first, actions are caused by an agent's desires and beliefs and, second, those desires and beliefs that cause action also rationalize the action. He then goes on to raise his principle criticism of this idea:

> "I think that the standard story is flawed in several respects. The flaw that will concern me in this paper is that the story fails to include an agent – or, more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but nobody – that is, no person – *does* anything. Psychological and physiological events take place inside

[102] J. David Velleman (1992): 'What happens when someone acts?', in: *Mind* (101), 461-481, 461.

> a person, but the person serves merely as the arena for these events: he takes no active part."[103]

Velleman criticizes that an action explanation that refers solely to mental states has only processes in mind that happen to take place within the agent, but which do not involve the agent as an agent who acts. If we ask ourselves, for example, whether or not Peter's yelling at Nancy was an action, causal action theorists would look at the etiology of this behavior to determine whether some mental states of Peter's brought this behavior about. Let us assume that the causal action theorists would come to the conclusion that Peter's yelling was an action, and, furthermore, that he gives the following explanation for having yelled: Peter yelled at Nancy because he had a desire to express his anger, he believed that yelling at her would do the job, and these mental states caused his behavior. Now, the challenge of the missing agent says that we don't know whether it was really a full-fledged action of Peter. After all, we only know that some desire of Peter – in concert with a particular belief of Peter – caused him to behave in a certain way. What we do not know, however, is how Peter was engaged in this behavior. The causal action theorist needs to answer that Peter's agency and his engagement in action is constituted or realized by the proper functioning of his mental states. The challenge, then, is to give an argument why we should believe that this is the case. In other words, why should the functioning of some proper parts of the agent constitute or realize her engagement in action?

If we cannot answer this challenge, this would indeed undermine the plausibility of a naturalistic account of agency and action. After all, what such a theory aspires to explain is agency, and it certainly fails to achieve this if it lacks the resources to account for the fact that agents perform actions. A central challenge for naturalistic accounts of agency, then, consists in showing that it does not eliminate the agent by restricting itself to an event-causal framework. And we should bear in mind that this challenge appears to be even stronger when we are concerned with more demanding forms of agency like autonomous agency since the autonomous agent is involved in an even more emphatic sense in her actions. I answer this challenge in Chapter 5.

[103] J. David Velleman (1992), 461.

3.4 Conclusion

The concept of autonomy under investigation conceives of autonomy as a natural property of persons. In this chapter, I have presented the ontological requirement that accounts agency and action in general, and autonomous agency in particular, have to fulfill in order to count as naturalistic ones. The basic requirement consists in staying within an event-causal ontology. Although some intuitions about the kind of control and independence that characterize autonomous agency give an initial plausibility to attempts that explain autonomy with reference to dualism, agent-causation, or volitions as irreducible mental actions, such non-natural ontologies contradict the idea of natural autonomy.

Now that we have formulated the naturalistic requirement, we need to investigate how an account of agency and action in general, and of autonomous agency in particular, can fulfill it. How can we account in event-causal terms for the specific control agents exert in action? Chapter 4 answers this question. There I address, first, the topic of agential control, and secondly, I answer the challenge of missing control. Against the background of a clearer notion of agential control, I then develop a causal account of agency and action. This account provides the action theoretical underpinning of the concept of autonomy under consideration.

4. Actions and Agential Control

In the last chapter, I presented the challenge of missing control. The challenge of missing control raises doubts about the possibility of giving an adequate account of agential control within a naturalistic framework. It says that control falls out of the picture if we restrict ourselves to an event-causal ontology. If this were true, we would indeed have a serious problem because actions are essentially characterized by control. Actions are exercises of control, or so I will argue in this chapter. Hence, without agential control, there is no action. And without a naturalistic account of agency and action, the whole idea of natural autonomy would be in jeopardy because autonomous agency is marked by using one's abilities to act in a particular way, that is, by making a special use of one's agency. If we cannot refute the challenge of missing control, we will fail to account for natural autonomy.

Let me highlight the problem again as it presents itself to us when we try to explicate the notion of autonomy. In Chapters 1 and 2, I pointed out that autonomy concerns the agent's dispositions and abilities to authentically express herself against opposition. Actions take center stage in the concept of autonomy under consideration because only through her actions can an agent actively express what is important to her. We shape our lives through our actions. If an agent is unable to act, she lacks the foundation on which autonomy rests because she lacks the ability that allows her to express what is important to her. For this reason, the explication of the notion of natural autonomy requires us to have a naturalistic account of agency and action. And the challenge of missing control is directed against such a naturalistic understanding of agency and action. For this reason, it also threatens the attempt to understand autonomy within a natural ontology.

I start this chapter with an explication of the notion of control. After all, in order to examine the claim that actions are exercises of control, we need to know what I mean with control. This also will help to get a clearer grip on the challenge of missing control. After an explication of the notion of control in section 4.1, I will argue for the claim that actions are exercises of control. (4.2) In particular, an agent needs to exercise rational control over her behavior. Apart from its intuitive appeal, this understanding of actions is backed up by the idea that actions are essentially goal-directed and done for a reason. I continue by explicating how we can account for actions as exercises of control within an event-causal framework. After all, this is

what needs to be done in order to refute the challenge of missing control. As will become apparent, the mental state of intention takes center stage in a naturalistic understanding of agency and action. I discuss the functional characteristics of intentions and how they are causally involved in action. (4.3-4.4) I conclude this chapter by introducing what I call the structured cause account of intentions. (4.5)

4.1 The Notion of Control

In order to make sense of the idea that actions are exercises of control, I first need to explicate what I mean by control. Here is how I – following Michael Pauen[104] – conceive of it. Control is a relational property that holds between someone or something that has control and someone or something that is controlled. In other words, we have a subject and an object of control. Following everyday ascriptions of control, typical objects of control are events, processes, persons, and states of affairs. I shall not systematically investigate what kinds of entities can be the objects of control, but I suppose that it all boils down to the control over events and states of affairs. The subject of control is typically taken to be some sort of agent. However, we can also think of natural or artificial systems, like thermostats, as controllers. I don't intend to systematically engage in disentangling the differences in all these different types of controllers. My focus lies on the necessary features of the control relation. Before I discuss them, though, I need to distinguish between *exercising* control and *having* control. A controller can have control over something without making use of it. I might be able to control what music is played on my stereo without actually doing anything regarding my stereo. A captain might be in control of the course the plane takes, although the plane is presently guided by autopilot. Now, if I pick a CD from my shelf, put it in my stereo, and start it, I am controlling what my stereo plays. And if the captain turns off the autopilot and grabs the steering wheel, she is exercising her control. When I speak in what follows about control, I have an ongoing process of control in mind, in contrast to the control that someone or something might have without exercising it.

[104] Michael Pauen analyzes the notion of control in his book *Illusion Freiheit?*. He highlights the three aspects of control that I view as being constitutive of what I call robust control: goal-directedness, causal influence, and monitoring. Compare Michael Pauen (2004): *Illusion Freiheit? Mögliche und unmögliche Konsequenzen der Hirnforschung* (Frankfurt a. M.: S. Fischer Verlag).

Control as an unfolding process involves two constitutive aspects: first, the controlling system needs to represent a goal. Control is always directed towards a certain goal. Second, the controller must causally influence the object of control in a way that leads to the achievement of the goal (or at least significantly furthers the achievement of the goal). Taken together these two aspects constitute what I call *minimal* control. Minimal control develops into *robust* control if a third feature is present, namely, monitoring. Often the controller monitors its performance and the progress towards reaching the goal in order to readjust its influence. Monitoring enhances control. Let us discuss these three aspects in turn.

First, control is necessarily goal-directed. It is impossible to describe a control relation in which the controller lacks a goal. Imagine an entity *E* that has no *p*-directed goals. *E* is completely indifferent with respect to the occurrence or non-occurrence of *p*. This means, among other things, that *E* has no dispositions whatsoever to engage in any sort of behavior *b* as a means to bring about or to avoid that *p*. If the agent engages in some *b* that brings about *p*, this is merely coincidental. Hence, *E* is not controlling *p*. As I stated, it is possible that *E* behaves in a way that affects whether or not *p* is the case. However, this would only be a side effect. If Anna, for example, makes herself a cup of coffee and incidentally influences the room temperature, she is not controlling the room temperature. How could she, when she does not care at all about the minute changes in the room temperature that result from her making some coffee? If Paul accepts a friend's invitation to come over for a beer, Paul's behavior influences the amount of beer at his friend's home. However, it would be rather odd to claim that he is controlling the amount of beer at his friend's home because he is completely indifferent with respect to this state of affairs. If it were otherwise, the notion of control would collapse into the notion of influence. Without a goal regarding *p*, *E* is not controlling *p*, even if *E*'s behavior influences *p*. Goal-directedness is a requirement for control.

The second requirement for control is the ability to exert causal influence in a way that secures achieving the goal or that significantly raises the probability of achieving the goal. To see this, imagine that *E* has a goal with respect to *p*. Let's say *E* wants it to be the case that *p*. Is *E*'s having a goal sufficient for *E*'s having control over *p*? Certainly not. Malcolm desires to go on holiday in Egypt. Let us imagine that he has the goal of going on holiday in Egypt. However, he lacks the time, the money,

and other resources that would enable him to organize a trip to Egypt. In this scenario, Malcolm has no control whatsoever with respect to whether he can make this trip. Take another example: even though Karla has the goal of studying at Smartplace, a private university that selects its students on a strictly random basis, i.e., by tossing a coin, she has no control over her being accepted at this university since she lacks any means to influence the selection process. Malcolm and Karla are lacking control because they are unable to influence whether they achieve or don't achieve their goals. It is a necessary condition for control that the controlling system is able to exert causal influence in a way that helps to secure the achievement of the goal. In this context, the goal has a guiding function. It allows the controller to pick those kinds of behavior that facilitate goal-achievement. After all, different goals are achieved by different means and choosing adequate means for realizing one's goals is a necessary condition for successful control.

The first requirement of control, the goal-directedness, precludes mere causal influence from being an instance of control. The second requirement accounts for the fact that control is a causal relation. Together they allow us to characterize minimal control as a goal-directed exertion of influence that secures goal-achievement. Robust control, in contrast, exhibits a third feature of control, namely monitoring. Scenarios in which the control behavior needs to be adjusted to the context in order to reach the goal provide the strongest case for the necessity of monitoring. If the controller needs to be sensitive to the information about the environment in order to adjust her behavior accordingly, monitoring becomes necessary. For example, if Lionel Messi has the goal of shooting a goal, and part of his plan for doing this is playing a one-two with Xavi, he needs to constantly monitor his own movements as well as the movements of Xavi, the defenders, and the ball. The success of the one-two depends on a constant monitoring of the situation and an adequate adjustment of one's own behavior. Take another example: building a house. If I want to build a house, I need to be aware of the progress in order to decide on my subsequent moves. I also need to take into account such things as the weather or the availability of building material. And since I don't build it alone, I need to coordinate my actions with the actions of other people, which requires that I observe our interaction in order to adapt my behavior according to what I need to do next. Sometimes what I need to do in order to reach my goal is not specified in advance in such detail that I can just trigger a behavioral chain that unfolds blindly. In such cases, control depends on monitoring.

Monitoring is not needed if there is surefire way for a controller to bring about the intended goal. If I can raise my arm without doing anything else, there is no need for a monitoring device. I can just do it. The best examples for the need of monitoring are more complex actions that are part of a larger project. In such a case, the agent needs to monitor the progress towards her goal and the contextual changes in order to adjust her behavior accordingly.

Let me summarize what I have said so far about control: minimal control is the capacity to exert a causal influence over an entity in order to secure the achievement of a specified goal. In addition to the capacity for a goal-directed exertion of influence, some controlling systems possess a monitoring device that enables them to flexibly adjust their behavior, that is, they possess robust control. Robust control is necessary for flexible behavior. Human agency is characterized by control that also exploits a monitoring device. How can we, against this background, corroborate the claim that actions are exercises of control?

4.2 Actions as Exercises of Control

What are actions? A good starting point for answering this question consists in contemplating what Enç calls "the first problem of action theory."[105] This is the problem of how to distinguish between behavior and action. Before we deal with this problem, let me mention an even more fundamental distinction, namely, the distinction between what an agent does and what happens to her. Getting a sunburn, being hit by a car, or receiving a love letter are things that happen to an agent. Scratching one's nose, blinking with one's eyes, or driving a car are things an agent does. Not everything an agent does is an action of her. The first problem of action theory is to distinguish between those things an agent does that are actions and those which are not. A terminological clarification is in order. Philosophers often speak about behavior as the category which actions are a part of. According to this convention, actions are a special kind of behavior. However, since behavior usually refers to bodily movements this terminology excludes mental activities like silently counting to ten, imagining a pink elephant, or deliberating about the question where to go on holiday. Instead of behavior, one could speak of an agent's being active or an agent's doings. But these formulations have the disadvantage of having a quite

[105] Berent Enç (2003), 39.

artificial ring to them. In what follows, I will sometimes adopt the convention and speak of behavior, and sometimes I will speak of an agent's doings or her activities. In all these cases, I understand this notion in a broad way that includes mental activity.

Having said that, what distinguishes mere behavior from action? Why do we sometimes count what an agent does as an action and sometimes not? Reading the newspaper, ordering another cup of coffee, or driving a car, for example, are ways of being active that count as actions, whereas sneezing, blushing, or stumbling do not count as actions, although they are ways of being active. How can we explain this distinction between different ways of being active? A plausible answer to this question is that actions are in some distinct way under the control of the agent. Indeed, this seems a rather good description of our intuitions about actions. If I sneeze, I have no control over what I am doing. When I stumble and step on your foot, it would be wrong to say that stepping on your foot is an action of mine. Again we could explain this by highlighting the fact that I had no control over my stumbling. I think that John Bishop formulates a well-entrenched intuition when he says that actions are "exercises of control."[106]

The basic idea of actions as exercises of control is closely related to the idea that actions are done for a reason. A lot of philosophers claim that what distinguishes actions from other behavior is that they are done for a reason. Donald Davidson expresses sympathies for this idea when he models his account of a primary reason for action in a way that upholds "the possibility of defining an intentional action as one done for a reason."[107] The same holds for John Bishop who points out that "[t]ypically (even essentially?) actions are 'done for reasons.'"[108] G. F. Schueler makes the even stronger claim that it seems to be "conceptually true" that "all actions are done for some reason."[109] It is a widely shared assumption in contemporary action theory that actions are done for a reason. This suggestion gives credit to the fact that actions can be typically explained by spelling out what the agent saw in so acting. When we say that actions are exercises of control, we account for this intuition. What is the connection between control and acting for a reason? Having control is a necessary

[106] John Bishop (1989), 23.
[107] Donald Davidson (1980 a): 'Actions, Reasons, and Causes', in: Donald Davidson (1980): *Essays about Actions and Events* (Oxford: Oxford University Press), 3-19, 6.
[108] John Bishop (1989), 99.
[109] G. F. Schueler (2003): *Reasons & Purposes. Human Rationality and the Teleological Explanation of Action* (Oxford: Oxford University Press), 1, FN 1.

condition for doing something for a reason. If I were lacking control over my behavior, it might happen that I do what I have reason to do, but only as a matter of coincidence. I would not do it because I have this reason.

The claim that actions are exercises of control is also supported by the observation that actions are essentially goal-directed.[110] We are purposive agents. That an action is goal-directed means that the agent's goal partly explains the agent's behavior. The behavior is a means of achieving the goal.[111] In other words, actions are goal-directed exertions of influence. And this is exactly the control relation we just spelled out. Hence, the fact that actions are goal-directed makes it true that actions are exercises of control.

An agent has a goal if and only if she represents a state of affairs as something that she wants to bring about.[112] Not every representation is a goal. For that, it is necessary that it be represented in the right mode. For goals, it is specific that they have a direct relevance for action as being end-setting and motivating. Let me say a little bit more about their end-setting function.

A goal can either be external to the action (if it is a consequence of the action), or it can be itself an action. If my goal is that my house is painted red, then my action of painting the house is apt because one of its consequences is that I achieve my goal. When my goal is to dance the tango with the beautiful dancer from Argentina, then my action of dancing with her is in and of itself enough for goal achievement. Goals can be more or less specific. I can want to make holidays at some point in the future (a rather unspecific goal), or I can want to fly for two weeks to Spain in August (a much more specific goal). A specific goal determines specific actions for achieving it. An abstract, less concrete goal needs to be specified before I can derive specific actions from it. If I have decided to go on holiday in Spain, I can make my travel arrangements. If I just know that I want to go on holiday but have not yet decided where and when, I need to specify my goal in more detail. Normally, we have a lot of goals at different levels of specificity. Some goals are more distal than others. If I want to drink a coffee right now, my goal is directed towards the immediate future. If

[110] Compare Alvin Goldman (1971): *A Theory of Human Action* (New Jersey: Prentice Hall) and Michael Smith (1987): 'The Humean Theory of Motivation', in: *Mind* (96), 36-61.
[111] This needs to be understood in a broad way, including constitutive relations and intrinsic goals.
[112] I use the notion of a state of affairs in a broad sense, which includes events. Compare D. M. Armstrong (1997): *A world of states of affairs* (Cambridge: Cambridge University Press).

I plan to drink a glass of wine once I am done with my tax report, my goal lies in the future.

Goals give us a starting point for planning specific actions. If I have a goal, I represent a certain state of affairs as one that I want to bring about. It is possible that I need the help or participation of others to bring this state of affairs about. But if it is really a goal of mine, I need to be able to have an impact on whether or not the state of affairs is realized. Hence, it cannot be a goal of mine that Brazil wins the World Championship since I have no influence whatsoever on bringing this about. For the same reason, I cannot have goals concerning the past. If I represent a state of affairs as a goal, I implicitly represent myself as being able to do something that makes a realization of this state of affairs more likely. This connection between goals and actions, which are conceived of as being conducive for realizing the goal, grounds the practical relevance of goals. If an agent adopts a goal, she has to ask herself what she can do to achieve this goal. Having a goal provides a direct input for practical deliberation. Projects and plans are, at bottom, goals combined with action steps for achieving the goal.

One consequence of the observation that actions necessarily are goal-directed is that they can be explained teleologically. We can explain why someone performed a certain action by pointing out that she had the goal of performing this action. Again, this presupposes that the agent has control. Hence, we can see that actions are basically exercises of control when we acknowledge that actions are goal-directed and done for a reason. To describe this kind of agential control, let me introduce the notion of *rational control*.

An agent possesses rational control over her behavior if and only if she represents a state of affairs as a goal and behaves with the aim of realizing this goal because she is in this representational state. In other words, mental states of the agent, for example, her desires and beliefs, represent a state of affairs as a goal and suitably cause the agent to perform certain behaviors as a means for realizing this state of affairs. These mental states constitute the agent's *motivating reason*.[113] If such a motivating reason is operative, the agent exerts rational control. The next section

[113] Compare for the notion of a motivating reason Michael Smith (1987); Philipp Pettit (1987): 'Humeans, Anti-Humeans, and Motivation', in: *Mind* (96), 530-533; Jay Wallace (1990): 'How to argue about practical reason', in: *Mind* (99), 355-385; Derek Parfit/John Broome (1997): 'Reasons and Motivation', in: *Proceedings of the Aristotelian Society. Supplementary Volumes* (71), 99-146.

deals with the issue of rational control and its application in a naturalistic account of agency and action in more detail.

4.3 Rational Control and Causal Etiology of Actions

Actions are exercises of control. In particular, the agent exerts rational control, that is, she controls her behavior based on her motivating reasons. In this section, I discuss in more detail how rational control is actually realized in agents like us. In effect, I present a naturalistic account of agency and action. Naturalistic accounts of agency and action are also called *causal* theories of actions because they try to establish a notion of action according to which an action is a piece of behavior with a particular causal etiology.[114] How can we explicate a causal theory of action?

Let us start with Donald Davidson who influenced virtually all contemporary causal theories of action with his seminal essay "Actions, reasons, and causes."[115] We can view him as the modern ancestor of causal theories of action. He starts out by asking the question: "What is the relation between a reason and an action when the reason explains the action by giving the agent's reasons for doing what he did?"[116] His answer, in brief, is that a reason "*rationalizes* the action," whereby "rationalization is a species of causal explanation."[117] In order to spell out this idea, he introduces the notion of a "primary reason."[118] According to Davidson, primary reasons figure as the causal mental antecedent of action.

As I explicated above, actions are explainable by citing the agent's reasons. This is just another way of saying that actions are under rational control. And reason explanations are always teleological explanations because they refer either implicitly or explicitly to the agent's goal in performing that action.

Davidson argues that in order to be successful, a reason explanation needs to imply that an agent acts because of the interplay of two kinds of mental states that can broadly be characterized as a conative or motivational mental state on the one hand,

[114] Bishop puts it this way: "To put it briefly, according to this Causal Theory of Action, to act is to be caused to behave by mental states of one's own – mental states that make the behavior reasonable in the circumstances." John Bishop (1989), 10.
[115] Donald Davidson (1980 a).
[116] Donald Davidson (1980 a), 3.
[117] Donald Davidson (1980 a), 3.
[118] Donald Davidson (1980 a), 4.

and a doxastic or epistemic mental state on the other. "Whenever someone does something for a reason, therefore, he can be characterized as (a) having some sort of pro attitude toward actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind."[119] Together, these two mental states cause the action or, more precisely, the event that is intrinsic to the action. Davidson calls such a pairing of a pro-attitude and a belief a "primary reason" and uses this notion in the formulation of his two central claims:

> "1. In order to understand how a reason of any kind rationalizes an action it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason.
>
> 2. The primary reason for an action is its cause."[120]

Pro-attitudes are sometimes conceived of as desires. Let us be clear, however, that Davidson uses the notion of a pro-attitude in a very broad sense. He points out that the category of pro-attitudes includes, "desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values in so far as these can be interpreted as attitudes of an agent directed toward actions of a certain kind."[121] The unifying feature of all these pro-attitudes is that they either are or imply a motivational force to perform an action of a certain kind. What is certainly not implied is that all these pro-attitudes share a feature that is often attributed to desires, namely that the agent feels good or satisfied when she gets what she desires and that this anticipated satisfaction is part of the agent's motivation for action.[122] Since Davidson does not equate pro attitudes with desires in our everyday understanding of desire, we need to be cautious to count him among the proponents of a Humean desire-belief model.[123]

---

[119] Donald Davidson (1980 a), 3 f.
[120] Donald Davidson (1980 a), 3.
[121] Donald Davidson (1980 a), 4.
[122] An excellent discussion of the notion of desire and its different usages is G. F. Schueler (1995): *Desire. Its role in Practical Reason and the Explanation of Action* (Cambridge MA: MIT Press).
[123] A Humean desire-belief model views desires as the only motivational states. For an extensive discussion of Hume's approach towards motivation, compare Francis Snare (1991): *Morals, Motivation and Convention* (Cambridge: Cambridge University Press). A quite influential formulation of a contemporary Humean model is Michael Smith (1987).

The belief-component of a primary reason says that a certain kind of action is apt to satisfy a desire, or pro-attitude, of the agent. Here is Davidson's more systematic definition of a primary reason: "*R* is a primary reason why an agent performed the action *A* under the description *d* only if *R* consists of a pro attitude of the agent towards actions with a certain property, and a belief of the agent that *A*, under the description *d*, has that property."[124] This characterization of the belief-component is based on the consideration that actions are intentional under a certain description. Hence, an action explanation in terms of an agent's primary reason must pick out the belief of the agent that represents the action as having the desired property. If I give you a present because I care about you and think that you will like the present, it would be wrong to say that I give you the present because I want you to express gratitude, even though I might also believe that you will express gratitude when I give you the present. If I did not settle for this action, i.e., giving you a present, because I believed that it will lead you to express gratitude, then it is false to explain my action by citing this belief as part of my primary reason. I did the action under a different description, namely, because I wanted to make you happy. Here is another example: a doctor reanimates a person, but in the process breaks some of her ribs. It would be wrong to say that the doctor acted as she did because she wanted to break the person's ribs. Even though she might have believed that her attempts to reanimate the patient might lead to some broken ribs, this was not what she wanted. The action appealed to her under the description of being a reanimation and not of being a breaking of the ribs. One and the same action can be described in hundreds of ways. But not all of these descriptions can be used in stating the agent's primary reason for doing what she did. For this, we need to find out under what description the action was intentional for her.

Equipped with the notion of a primary reason, we can formulate the foundation of a causal theory of action: an action is behavior that is caused by a

[124] Donald Davidson (1980 a), 5. What kind of properties of an action does Davidson refer to? He gives us no direct answer to this, but his examples and systematical considerations make it plausible that there are three kinds of properties that an agent could be interested in. First, an action can be conceived of as a *means* for achieving some goal. That is, actions can have this kind of instrumental value. Buying a bottle of wine is a means for having a nice dinner, for example, at least when you like a good glass of wine with your pasta. Second, an action can be a *constitutive part* of a goal of an agent. For example, playing with one's children is a constitutive part of being a good parent. Giving a good sermon is a part of holding a good service. Third, an action can be conceived of as being *intrinsically attractive*. Dancing a tango, for example, or doing what's morally required, can be in itself a goal of the agent.

primary reason. Primary reasons provide us with a teleological explanation because they represent performing a certain kind of action as a goal. At the same time, the primary reason is, as Davidson argues, the cause of the action.

What is Davidson's argument for assuming that primary reasons are causes? He asks us to imagine an agent who has the mental states that constitute a primary reason for action, and who performs the very kind of action that the primary reason specifies. Davidson assumes that we can say that this action was reasonable in light of her mental states. But this alone does not suffice to explain her action. If we only mention that the person has these mental states, "something essential has certainly been left out, for a person can have a reason for action, and perform the action, and yet this reason not be the reason why he did it. [sic] Central to the relation between a reason and an action it explains is the idea that the agent performed the action *because* he had the reason."[125] And he continues that our best interpretation of this 'because' is a causal one.

Alfred Mele illustrates this idea rather elegantly when he writes:

> "Goldman and Davidson agree that an action A is done for a reason R only if R is a cause of A. Here they are on firm ground. Arnold has a reason for leaving the lecture hall: He wants to display his dissatisfaction with the lecturer's sexist remarks and believes that leaving the room is a means of doing so. If he does leave the room, do we have here a sufficient condition of his having done so *for* the reason just identified? Plainly not. He might have left the room for another reason altogether. Perhaps he recalled an important dental appointment and left the lecture in order to catch a bus to the dentist's office. The reason for which he leaves the room is, as we might say, the reason that *accounts for* his leaving the room. And it is difficult to see how a reason can account for someone's A-ing if it (or the agent's *having* it) does not play a suitable role in the etiology of his A-ing."[126]

[125] Donald Davidson (1980 a), 9.
[126] Alfred R. Mele (1992): *Springs of Action. Understanding Intentional Behavior* (Cambridge: Cambridge University Press), 7.

The challenge for the philosopher who denies that reasons are causes, then, is to explain how we can distinguish the reason for which an agent acts from those reasons that she has for performing this action but which are not operative if not by reference to their causal role.[127]

John Bishop develops the causal theory of action more systematically based on Davidson's account of primary reasons. The first formulation of the causal theory of action (CTA) is basically a reconstruction of Davidson's ideas:

> "CTA-H:
>
> M performs the intentional action of *a*-ing if, and only if,
>
> (1) *M* is in a complex mental state, *r*;
>
> (2) *M's* being in *r* makes it reasonable for *M* to do *a*;
>
> (3) *M's* being in *r* causes an outcome, *b*; and
>
> (4) *b* instantiates the type of state or event intrinsic to the action of *a*-ing."[128]

(1) and (2) are meant to capture the teleological nature of action explanations, that is, the fact that a-ing is a goal of the agent. (3) and (4) secure that the agent's mental states indeed explain the action because they cause the behavior. We can say that, according to CTA-H, rational control is realized by the complex mental state r.

As it stands, this initial formulation of a causal theory of action is open to criticism because it fails to answer the challenge from deviant causal chains. In order to deal with these problems, we need to specify both the mental states and the proper causal relations between the mental states and the behavior in more detail. However, this formulation already gives us a good outline of a causal theory of action. I will come back to the question of what mental states are required and how we need to spell out the causation so that it is not victim to deviance. My answer will be that intentions play the central role and that the behavior needs to be sensitive to the content of the

[127] At this point we should note that a similar problem arises for proponents of agent causation. If the agent is the ultimate source of an action, it is unclear how her reasons can explain her action.
[128] John Bishop (1989), 104.

intention. Before I come to these issues, I want to stay a moment longer with the topic of rationalization and reasonableness. Davidson says that reasons rationalize actions, and Bishop claims in (2) that the mental states that cause the behavior "make it reasonable" for the agent to perform the action. How shall we understand that?

Bishop emphasizes the necessity of rational control for action. He argues that we need to refer to rational control in order to explain why certain behavior is not an action although it is caused by the mental states of the agent. Imagine, for example, that I always start to shiver when I think of my math teacher. My shivering is caused by my thinking about my math teacher. Hence, we have mental causation. However, my shivering is not an action of mine because the connection between my mental states and my behavior, i.e., shivering, is not right. What is wrong with it? "Given the rationality condition, behavior cannot count as intentional action unless it is rational in the minimal sense that it is reasonable *with respect to the intentional states that cause it*."[129] My shivering then is not an action of mine because thinking about my math teacher is not a rationalizing cause of my shivering. In other words, my shivering is not under my rational control.

The notions of rationalization and reasonability that we use in this context are meant in a minimal and subjective sense. They do not refer to an all-things-considered judgment (hence minimal), and they do not refer to an objective standard of rationality, but rather to the agent's perspective (hence subjective). Doing something for a primary reason does not imply that the agent is overall and objectively rational. This understanding is in the background when Bishop says: "if a CTA analysis is correct, part of the concept of intentional action is the idea of behaving in a way made reasonable, relative to *the relevant* intentional states of the behaver, however, irrationally held these themselves may be."[130]

Let me introduce the distinction between motivating and normative reasons at this point.[131] Motivating reasons are those mental states that effectively motivate the agent in performing a particular action. The standard assumption is that motivating reasons are akin to what Davidson calls a primary reason, that is, motivating reasons

[129] John Bishop (1989), 105.
[130] John Bishop (1989), 111.
[131] For a good systematical discussion of these notions compare the references in FN 113; J. David Velleman (1996): 'The Possibility of Practical Reason', in: *Ethics* (106), 694-726; Jonathan Dancy (2000): *Practical Reality* (Oxford: Oxford University Press).

are pairs of desires and beliefs. In contrast to that, a normative reason is a consideration an agent takes as speaking in favor of a particular kind of action.[132] It is quite possible that an agent's normative reasons are not the reasons that move her to action. The notion of rational control as I have explicated it here does not imply that the agent acts for a normative reason. Behavior is under rational control if it is done for a motivating reason. As I argue in Section 5.3, normative reasons are necessary for determining an agent's own authentic standpoint. But for action as such, motivating reasons are sufficient.

In order to clarify the issue of rationality in action, Bishop contemplates akratic actions. In akratic action, the agent judges that she ought to A but B's instead, whereby B-ing is also an intentional action of the agent. Bishop explains that akratic action poses a problem for a causal theory of action because it is *prima facie* unclear how akratic action can be caused by mental states that rationalize the action. He points out that the following, initially plausible rationality condition is unable to account for akratic action: "If agents judge that it would be better to do x than to do y, and believe themselves free to do either x or y, then they will do x intentionally if they do either x or y intentionally."[133] According to this principle, it is a necessary condition for action that the agent acts in accordance with her overall judgment about what she ought to do. The possibility of akratic action, however, falsifies this principle because akratic action is defined as action against one's better judgment. "Akrasia, then, poses a problem for a CTA analysis because it excludes what might otherwise seem to be the most natural specification of the rationality condition for intentional action."[134]

According to Bishop, the possibility of akratic action forces us to develop another account of the rationality condition. "To defend a CTA analysis, an alternative specification of the rationality condition is required, according to which even akratic acts can count as relatively rational with respect to the defining mental causes of intentional action."[135] His solution is to introduce the notion of intention as that kind of mental state that makes actions, even akratic ones, "relatively rational." In

[132] Thomas Scanlon argues that the notion of a normative reason is not further reducible. "I will take the idea of a reason as primitive. Any attempt to explain what it is to be a reason for something seems to me to lead back to the same idea: a consideration that counts in favor of it. 'Counts in favor how?' one might ask. 'By providing a reason for it' seems to be the only answer." T. M. Scanlon (2000): *What We Owe to Each Other* (Cambridge MA: Harvard University Press), 17. I agree with Scanlon.
[133] John Bishop (1989), 111.
[134] John Bishop (1989), 111.
[135] John Bishop (1989), 111 f.

this he is again following Davidson, who uses the notion of an unconditional practical judgment in order to be able to deal with akrasia. An unconditional practical judgment just says that I ought to A, period. Based on this idea, we can formulate a new principle for the rationality requirement, namely: "If agents *judge unconditionally* that it is better to do x than to do y, and believe themselves free to do either x or y, then they will do x intentionally if they do either x or y intentionally."[136] And this principle gives us the necessary room to deal with akratic action. Against this background, we can understand every action to be "relatively rational" with respect to an unconditional practical judgment. And since it is possible that this judgment differs from the all-things-considered judgment about what one ought to do, we can act akratically.

Bishop himself acknowledges that "[t]he claim that there is a kind of practical judgment that is more 'all out' and unconditional than a final evaluation of what it is best to do, all things considered, is psychologically implausible. It is not easy to get an independent grip on what unconditional practical judgments really come to."[137] He then concedes that introducing unconditional practical judgments is *ad hoc*. In order to amend these worries, Bishop proposes that unconditional practical judgments should be understood as intentions. "The odor of ad hocness may be somewhat dispelled, however, once we recognize that Davidson's unconditional judgments are not strictly evaluative judgments at all but are really to be identified as *intentions to act*."[138] What Bishop suggests, then, is that every action is based on an intention. And intentions are not "evaluative judgments," that is, they are not necessarily expressive of an overall judgment. "If *r* is construed as necessarily consisting in an evaluative judgment in favor of doing *a*, or some set of mental states (beliefs and desires, for example) that themselves justify such a judgment, then CTA-H is falsified by the possibility of akratic actions. From now on, then, we shall understand CTA-H so that *r* can consist merely in a final stage intention."[139]

In order to highlight the minimal and subjective sense of rationality involved in this context, Bishop refers to the notion of a match between the representational content of the intention and the action:

[136] John Bishop (1989), 113.
[137] John Bishop (1989), 113.
[138] John Bishop (1989), 114.
[139] John Bishop (1989), 117.

> "According to CTA-H, for M to perform the intentional action of a-ing, M must be in a mental state, r, that makes the outcome, b, *reasonable* under the description 'a-ing.' For brevity, let us say that an outcome, b, which stands in this relation to a mental state, r, 'matches' r's content, and we shall refer to this requirement of a CTA analysis as its 'condition of match.'"[140]

The matching condition is essential for an adequate causal explanation of rational control because, as we will see, it solves the problem of deviant causal chains. Hence, the introduction of intentions is necessary in order to develop an adequate causal theory of action. The next section is devoted to intentions.

4.4 Intentions

The idea that intentions are a distinct kind of mental state that is not reducible to desires and beliefs is a relatively recent one in modern action theory. Until late into the second half of the 20th century, philosophers of action attempted to explain actions in terms of desires and beliefs – if they were not rejecting the idea of action explanation by reference to mental states altogether. Behaviorism was trying to get rid of mental explanations rather successfully for a while. But even after it became obvious that behavior was not adequately explainable solely in terms of stimulus-response patterns, it took a while till philosophers like Myles Brand[141], Michael Bratman[142], or Alfred Mele[143] rehabilitated intentions as a distinct kind of mental state. A parallel development took place in psychology where the concept of the will was largely in discredit due the legacy of behaviorism until people like Julius Kuhl[144],

[140] John Bishop (1989), 128.
[141] Myles Brand (1984): *Intending and Acting. Toward a Naturalized Action Theory* (Cambridge MA: MIT Press).
[142] Michael E. Bratman (1987): *Intention, Plans, and Practical Reason* (Cambridge MA: Harvard University Press).
[143] Alfred R. Mele (1992).
[144] Julius Kuhl (1985): 'Volitional Mediators of Cognition-Behavior Consistency: Self-Regulatory Processes and Action Versus State Orientation, in: Julius Kuhl/Jürgen Beckmann (eds.) (1985): *Action Control. From Cognition to Behavior* (Berlin: Springer-Verlag), 101-128. Compare also the historical overview of the psychology of action control, which also deals with the "neglect of volitional processes in psychological research." (Julius Kuhl/Jürgen Beckmann (1985): 'Historical Perspectives in the Study of Action Control', in: Julius Kuhl/Jürgen Beckmann (eds.) (1985), 89-100, 89.

Heinz Heckhausen, or Peter Gollwitzer[145] paved the way for a revival of a psychology of will that investigates processes of action control.[146] What are the arguments that force us to acknowledge intentions in our action theory?

The most important observation is that intentions carry a special kind of commitment towards a goal or an action. In contrast to merely desiring a particular course of action, intending to x settles the matter, at least for the moment. That is, intentions belong – just like desires – in the category of pro-attitudes.[147] In contrast to desires, however, intentions transport a special kind of commitment on the side of the agent. This commitment is a crucial difference to desires or other pro-attitudes. Michael Bratman investigates this commitment extensively. He distinguishes "two dimensions of commitment,"[148] a "volitional" one and a "reasoning-centred" one. Volitional commitment is concerned with action control.[149] Bratman conceives of intentions as mental states that directly control actions. A desire, in contrast, might or might not have an impact on what we do.[150] "Intentions are, whereas ordinary desires are not, conduct-*controlling* pro-attitudes. Ordinary desires, in contrast, are merely *potential influencers* of action."[151] Thus, intentions have the function of controlling behavior because of their volitional commitment.

Intentions embed a reasoning-centered commitment because once an intention is formed, it is usually retained if the agent does not encounter relevant new

[145] Heinz Heckhausen/Peter Gollwitzer (1987): 'Thought contents and cognitive functioning in motivational versus volitional states of mind', in: *Motivation and Emotion* (11), 101-120.
[146] Recently Patrick Haggard contributed to a clearer understanding of intentions from a neuroscientific and psychological perspective. Compare Marcel Brass/Patrick Haggard (2008): 'The What, When, Whether Model of Intentional Action', in: *The Neuroscientist* (14), 319-325; Patrick Haggard (2008): 'Human volition: towards a neuroscience of will', in: *Nature Reviews Neuroscience* (9), 934-946. In collaboration with the philosopher Elisabeth Pacherie Haggard also reflects the philosophical implications of his work: Elisabeth Pacherie/Patrick Haggard (2010): 'What are Intentions?', in: Walter Sinnott-Armstrong/Lynn Nadel (2010): *Conscious Will and Responsibility* (Oxford: Oxford University Press), 70-84. Elisabeth Pacherie presented her own accounted of intentions in series of publications. Compare in particular Elisabeth Pacherie (2000): 'The Content of Intentions', in: *Mind & Language* (15), 400-432.
[147] "Both intentions and desires are, but ordinary beliefs are not, *pro-attitudes*. Pro-attitudes in this very general sense play a motivational role; in concert with our beliefs they can move us to act." Michael R. Bratman (1987), 15.
[148] Michael R. Bratman (1987), 15.
[149] "The first concerns the relation between intention and action, and I will call it the *volitional* dimension of commitment (or, for short, volitional commitment)." Michael R. Bratman (1987), 15.
[150] Of course, there are situations in which an intention does not lead to action, think of akrasia, for example. However, it is part of the function of intentions to guide actions. Having a desire does not rationally require that the agent tries to satisfy it. There is nothing wrong in not acting on certain desires. However, holding an intention and not acting on it is rationally criticizable.
[151] Michael R. Bratman (1987), 16.

information.[152] "Retention of my prior intention and nonreconsideration is, so to speak, the 'default option.'"[153] Because of this typical stability, intentions often play a crucial role in deliberation as Bratman points out: "I will frequently reason from such a prior intention to further intentions. I will frequently reason from intended end to intended means or preliminary steps [...] And I will frequently reason from more general to more specific intentions."[154] Bratman summarizes the essential characteristics of intentions as follows:

> "We have identified three kinds of dispositions associated with such states [intentions]. They are conduct-controlling – not merely potentially conduct-influencing – pro-attitudes; they resist reconsideration, and so have a characteristic inertia; and they play characteristic roles as inputs into further practical reasoning to yet further intentions."[155]

Being committed is the central feature of intentions.[156] Of course, it is possible to analyze the functional characteristics of intentions in greater detail. This is what I will do next.

Intentions are causes of actions. This is already implied in the observation that intentions are controlling attitudes. Alfred Mele certainly is right in claiming that it is common to regard intentions as initiators of action. He says that "[t]here is a widespread agreement among philosophers of action that intention is a motivating cause of intentional actions."[157] Berent Enç, for example, expresses this idea when he maintains, "This intention state ultimately causes the basic act at the appropriate time. Thus the intention is what eventually triggers the basic act."[158] The very same idea also underlies John Bishop's causal theory of action: "In general, the mental state that

---

[152] "My intention resists reconsideration: it has a characteristic *stability* or *intertia*." Michael R. Bratman (1987), 16.
[153] Michael R. Bratman (1987), 17.
[154] Michael R. Bratman (1987), 17.
[155] Michael R. Bratman (1987), 22.
[156] The focus on the special kind of commitment that characterizes intentions also plays a crucial role in Heinz Heckhausen and Peter Gollwitzer's psychological 'Rubikon'-model. Compare Heinz Heckhausen/Peter Gollwitzer (1987).
[157] Alfred R. Mele (1992), 130.
[158] Berent Enç (2003), 181. However, in a footnote he also says: "It is possible to assign the function of triggering the basic act to a separate module." (181, FN 4) It is unclear what Enç wants to claim here.

initiates action will be the agent's *intending* a certain goal or end, *e*."[159] Mele himself follows this understanding when he claims, "Effective intentions motivate intentional actions."[160] I shall assume that intentions either are motivational states or that they modify the motivational states of the agents.

A second functional characteristic of intentions is that they sustain actions. I agree with Mele when he says, "In addition to any triggering or *initiating* function that intentions may have in the etiology of intentional action, they also have a *sustaining* function."[161] An agent would not finish her action if she lost her corresponding intention. If I cease to intend to paint my room, I stop painting it. If you lose your intention to read this discussion, you stop reading it. Mele interprets the fact that ceasing to intend to A leads the agent to stop A-ing as speaking in favor of the idea that intentions incorporate motivation. "This indicates that the sustaining function of intentions […] is at least partly conative or *motivational*."[162] But again, the alternative interpretation I just mentioned is also possible. That is, intentions might sustain action by continuously influencing the agent's motivations without being a part of them.

A third functional characteristic of intentions is that they guide actions. Guidance consists in such things as determining which course of action to take and how to react to situational requirements. It is also a part of the guiding function to determine when to start acting and when to finish. In the case of complex actions or sequences of actions, intentions coordinate the interplay of the different parts of a complex action or, in the case of an action sequence, intentions help one to select the right action at the right time.

The guidance function depends on the representational content of intentions.[163] Intentions represent a goal and also what can be called an action plan. In the minimal case, the goal is a basic action of the agent, and the plan component consists just in a representation of performing this basic action. "If [an agent] intends an end that he or she can directly achieve, [the agent] already possesses a trivial action-plan for fulfilling such a 'basic' intention."[164] In contrast to the basic intention, we often form

[159] John Bishop (1989), 128.
[160] Alfred R. Mele (1992), 142.
[161] Alfred R. Mele (1992), 130.
[162] Alfred R. Mele (1992), 131.
[163] "The guiding function of intention rests on its plan component. An intention-embedded plan provides action directions." Alfred R. Mele (1992), 145.
[164] John Bishop (1989), 129.

what Bishop calls "non-basic intention."[165] What is special about non-basic intentions is that they represent actions that are instrumental to achieving the goal. "Agents develop action plans […] to discover what basic actions (directly under their control) they should take in order to achieve ends not directly under their control."[166] From a functional point of view, then, plans are necessary to successfully achieve one's end.

The fact that intentions are goal-directed and that they represent a plan of how to achieve this goal helps to guide actions since it helps to select the right action at the right time and it allows deciding whether the ongoing action has to be readjusted in order to achieve the goal. In addition, the fact that the success conditions of actions are represented not only allows for adapting to situational requirements, but also helps to decide whether one has to go on with A-ing or not.

Successful guidance depends on monitoring the ongoing action, situational changes, and the progress towards reaching the goal. Flexibility in performing actions depends on getting feedback about the current state of the process. New information has to be taken into account in order to successfully adapt one's own actions to situational requirements. Mele is right when he says, "In executing the intention, I am guided by the plan. This guidance is dependent upon the *monitoring* of progress toward my goal."[167]

Monitoring is sometimes regarded as a function of intentions. Myles Brand, for example, says, "The cognitive component of immediate intention is the guidance and monitoring of ongoing activity."[168] Mele himself leaves it open whether monitoring is among the functions of intentions. I shall simply assume that intentions exploit a monitoring system. I leave it open whether the monitoring system has to be understood as being a part of the intention or whether it is a separate module.

As I have already pointed out, some authors, most notably Michael Bratman and Alfred Mele, maintain that intentions also exhibit deliberative functions. It is indeed the case that intentions sometimes provide input for deliberations. And having an intention to A makes it redundant to engage in further deliberation about whether

[165] John Bishop (1989), 128.
[166] John Bishop (1989), 130.
[167] Alfred R. Mele (1992), 136.
[168] Myles Brand (1984), 173. Recently Jesus Aguilar and Andrei Buckareff have listed monitoring among the functions of intentions: "[I]ntentions are taken to be the executive states that coordinate, cause, sustain, guide, and monitor intentional behavior." Jesus H. Aguilar/Andrei A. Buckareff (2009): 'Agency, Consciousness, and Executive Control', in: *Philosophia* (37), 21-30, 25.

to A or not, as long as there is no new information which is regarded as relevant for this question. Thus, acquiring an intention to A often blocks deliberation about whether to do A, since the agent sees this question as settled. Mele describes the deliberative functions as follows: "intentions are plausibly regarded both as providing motivation to engage in practical reasoning with a view to their execution […] and as being well suited to put a proper end to practical reasoning. Some intentions […] prompt practical reasoning, while others *appropriately terminate* it."[169]

Intentions are also quite useful for the coordination of actions. Mele, for example, says that intentions "help coordinate agents' behavior over time and their interaction with other agents […]."[170] This coordinative function is based on the commitment that is characteristic of intentions. In contrast to desiring to A, intending to A establishes a special commitment towards A-ing. This, in turn, explains why, as Bratman puts it, "an intention to A normally supports a belief that the agent will A. And this belief helps to facilitate coordination."[171]

To summarize: intentions are mental states that initiate, sustain, and guide actions. Guidance requires monitoring. These functional characteristics of intentions make it the case that intentions are mental states that realize rational control.[172] Let us take a closer look at the matching condition and how we can explicate it with reference to intentions.

The potential matching between an intention and an action is due to the plan component of the intention. Here is Bishop's more detailed account of this matching condition:

> "For an outcome to match a given action-plan that passes this test, at least the following conditions are necessary. First, the agent must perform the basic acts specified in the plan. Second, the agent's doing so must be intentional under the description given in the plan. And, third, the agent must perform these acts with

[169] Alfred R. Mele (1992), 138.
[170] Alfred R. Mele (1992), 140.
[171] Michael R. Bratman (1987), 18.
[172] Borrowing a notion from Fisher and Ravizza, I would say that intentions realize "guidance control." An agent exerts guidance control, in contrast to alternative possibilities control, if she is in the right way causally involved in performing that action. Fisher and Ravizza argue that "an agent exhibits guidance control of an action insofar as the mechanism that actually issues in the action is his own, reasons-responsive mechanism." John Martin Fisher/Mark Ravizza (1998), 31.

> the intention of attaining the end the plan was formed to serve. […] a fourth condition of match with the action-plan should be that the actual outcome must conform to the agent's beliefs (formed in his or her practical reasoning) about how it would be that the basic actions planned would yield the desired goal."[173]

The first three conditions are necessary for performing an action. The fourth condition is important for deciding what action the agent actually performed. It is concerned with the outcome of the agent's action that is not directly under her control. Daniel Bennett gives the example of a killer who intends to kill his victim by shooting him.[174] The killer is performing the basic acts specified in the action plan but misses his target. However, the noise of the shooting scares a herd of pigs, which flees in panic, thereby trampling and inadvertently killing the supposed victim of the shooting. Because of the fourth condition, it is wrong to ascribe the action of killing to the killer. After all, his plan was to kill him with a bullet and not with some panic-stricken pigs. But it is out of the question that he performed an action, namely, the action of shooting at the victim. That his plan failed in some detail does not make it the case that he didn't perform any action at all. This is different when one of the first three conditions is violated. When I do not even perform a basic act, I am not acting at all. And when I perform the movements of a basic act but did not intend to perform it under a certain description, I am not acting.

The matching condition does not only help us to understand how the relation between mental states and behavior has to be in order to constitute action. As we have just seen, it is also an essential tool to deal with the problem of causal deviance. Deviant causal chains are a serious challenge for causal theories of action because they raise doubts about the possibility to account for action in purely event-causal terms. In examples of deviant causal chains, an agent's behavior is caused by mental states that also represent the outcome as something to bring about. Roderick Chisholm gives the example of the nephew who wants to kill his uncle in order to inherit the uncle's money.[175] The nephew desires to kill the uncle, and he believes that he can do so by shooting him. Let us assume that he forms the intention to kill his uncle by

[173] John Bishop (1989), 131 f.
[174] Donald Davidson discusses this example and credits Bennett as the one who developed it. See Donald Davidson (1980 b): 'Freedom to Act', in: Donald Davidson (1980), 63-82, 78.
[175] Roderick Chisholm (1966).

shooting him. On the way to his uncle's, the nephew is so agitated by his plan to kill his uncle that he drives recklessly. As a consequence, he runs over a pedestrian who dies. As it turns out, the pedestrian was his uncle. Hence, the nephew killed his uncle. Moreover, his intention was the cause of this killing because it made him drive recklessly. Here we have a match between the intended goal and the outcome of the nephew's action. But it is wrong to say that the nephew murdered his uncle. The challenge for a causal theory of action is to explain why this is so, even though the mental states that caused the agent to behave in a way that lead to the killing of his uncle were aiming at exactly this outcome.

The matching condition allows us to deal with some of these cases of non-basic causal deviance. In cases of non-basic causal deviance, the agent performs a basic act that is part of her action-plan, but this basic act brings about the intended outcome in a deviant way. I have already mentioned Bennett's example of the killer who inadvertently kills his victim by scaring a herd of pigs into a frenzied flight. Chisholm's murderous nephew is another example of this sort of non-basic deviance. With the matching condition at hand, we can also answer this kind of cases. The nephew was not murdering his uncle because running over an anonymous pedestrian was not part of his plan.

Unfortunately there is another form of causal deviance that the matching condition does not seem to be able to deal with. In basic causal deviance, the deviant causal chain does not occur between the behavior and the outcome, but between the intention and the behavior. A famous example for a basic deviant causal chain comes from Davidson. Davidson invites us to imagine a climber who is in a precarious situation: he is connected with a rope to another climber who has slipped and is now dangling in the air.[176] The climber who holds the rope is too weak to hold him much longer. He forms the desire to get rid of the weight by letting his companion fall. He also has the belief that he just needs to loosen his grip to achieve this. As a consequence of this horrifying desire, he becomes so agitated that he starts to tremble and let go of the rope. The climber achieved what he desired, and he achieved it as a consequence of behavior that was caused by the very mental states that were aiming at this outcome. Nonetheless, the trembling and the ensuing loosening of his grip appear not to be attributable to him as action. But why not? Bishop correctly observes that we

[176] Donald Davidson (1980 b).

want to say that the climber was not in control of losing his grip. Unfortunately, this answer is not open to a causal theory of action because reference to an agent's exercise of control would make the theory circular. What we need, then, is an explanation that accounts for the correct intuition that the agent is not control in an event-causal framework.

Bishop suggests what he calls a sensitivity strategy in order to deal with basic deviance. The central idea of this sensitivity strategy exploits the insight that intentions do not only specify a goal but also a plan, and their causal influence extends over the whole period of behavior. "The *sensitivity strategy*, then, suggests that a CTA analysis will exclude basic deviance if it includes the requirement that the caused behavior shows a certain responsiveness or sensitivity to the content of the intention that causes it."[177] In other words, the behavior that is intrinsic to an action needs to unfold in response to the plan component of the intention. Otherwise the behavior does not count as an action.

Berent Enç employs a similar idea when he suggests what he calls an "explanatory relation requirement"[178] in order to explicate the suitability condition and thereby deal with causal deviance. The definition of his causal theory of action, then, is as follows:

> "CTA: The behavioral output of an organism is an intentional action A if it is caused in the way it is supposed to be caused by an intention to do A.
>
> E_0: An intention to do A causes an event in the way it is supposed to if and only if for any intermediate link, X, from the intention to the event, the fact that the intention causes X is explained by the fact that X results in that event."[179]

Enç claims that this suggestion gives us a unitary treatment of deviant cases because basic and non-basic deviance both are characterized by causal chains that violate E_0. Take Davidson's climber. He forms the intention to let go of his companion by losing his grip on the rope. This intention disturbs him so much that he becomes shaky

[177] John Bishop (1989), 148.
[178] Berent Enç (2003), 111.
[179] Berent Enç (2003), 112.

which lets him lose his grip. That his intention causes him to be shaky cannot be explained by the fact that shakiness causes him to let go of the rope. Hence, the explanatory relation requirement is violated. The same is true for examples of non-basic deviance. The killing of the victim is not an intentional action of Bennett's shooter because the intention (via the action) caused the herd of pigs to run down the victim and cannot be explained by the fact that the pigs kill the victim. Again, the explanatory relation requirement is violated.

Enç exploits the assumption that intentions have a specific function, namely to cause the behavior they represent, thereby initiating a particular causal sequence that leads to the achievement of the agent's goal. The intention functions as it is supposed to if and only if it causes events because they finally lead to the causation of whatever goal the agent wishes to accomplish with this particular intention. It is always possible that the intention causes the desired outcome by fluke. However, in these cases we cannot explain the causal relation between the intention and the fluke events by pointing out that these events eventually lead to the desired outcome. In this way, they are not suitable for causing the behavior or the consequences of the behavior.

Enç does not deny that there is a difference between basic and non-basic deviance. All he claims is that we do not need distinct accounts for dealing with both of these cases.

> "When we examine the causal pathway in the framework of the function of the system, we can distinguish cases in which the system malfunctions (the category of antecedentially deviant cases) from those in which the system functions well, but the co-operation if the environment is totally fortuitous (the category of consequentially deviant cases). […] the requirement of explanatory relation provides a unified account not only for these two types of deviance, but other categories that have been since discussed."[180]

Examples for additional forms of waywardness include cases in which an agent achieves what she wants to achieve because she formed a false belief that led her to

[180] Berent Enç (2003), 113.

succeed in her action or in which an agent achieves her goal by relying on a completely randomized mechanism, like winning the lottery.

We now have all the elements we need for explicating a naturalistic notion of agency and action, that is, a causal theory of action. Behavior counts as an action if and only if it is under rational control. Rational control is realized by intentions. This means, in causal terms, that behavior is an action if and only if it is caused in the right way by the agent's intentions. The problem of causal deviance can be solved by introducing a sensitivity requirement, according to which the behavior has to be sensitive to the content of the intention. In the final section of this chapter, I want to discuss the question how intentions cause actions.

4.5 The Structured Cause Account of Intentions

In this chapter, I have developed a causal understanding of action. The mental state of intention stays in the center of this account, which says that actions are behaviors that are caused in the right way by the agent's intentions. I explicated how this 'in-the-right-way'-phrase can be understood by using the idea of sensitivity of behavior to the representational content of an intention. How do we have to understand the causal relation between intentions and actions? That is the question I want to discuss in this final section.

Intentions are causes of actions. I pointed this out in section 4.4. The majority of contemporary philosophers of action agree upon this. It is often left open how this causation has to be understood. I think it is fair to assume that most philosophers conceive of intentions as triggers of action. Myles Brand, for example, assumes that "there must be one type of event that is the proximate cause of action."[181] He calls this "proximate cause of action" an "immediate intention."[182] Similar ideas form the core of Michael Bratman's notion of a "present-directed intention"[183] and Alfred Mele's notion of a "proximal intention."[184] I want to raise some doubts about the idea that intentions are proximal causes of actions.[185] My aim is to argue for a different

[181] Myles Brand (1984), 35.
[182] Myles Brand (1984), 35.
[183] Michael R. Bratman (1987), 4.
[184] Alfred R. Mele (1992), 143.
[185] Let us grant for the sake of the argument that the notion of proximal causation makes sense in the first place.

understanding of intentions, which I dub the structuring cause account of intentions. Let me explain this in more detail.

The structuring cause account of intentions contrasts with what I call the triggering cause account. The former account views intentions as structuring causes of action, the latter views them as triggering causes. I borrow the distinction between triggering and structuring causes from Fred Dretske. Here is an example that should make it intuitively clear what general distinction I have in mind by speaking of triggering and structuring causes:

> A Jukebox plays a song every time someone types in a number between 1 and 100. The event of typing in a number is the triggering cause of the jukebox's playing a particular song. Imagine that typing in the number 50 causes the jukebox to play 'Billy Jean'. Now imagine that a technician changes the association between the numbers and the songs in a way that typing in a 50 does not cause 'Billy Jean' to be played, but 'Radio Gaga' instead. Under these circumstances someone's typing in a 50 is the triggering cause of 'Radio Gaga' being played by the jukebox. Now imagine that someone asks why typing in a 50 causes 'Radio Gaga' to be played. In this case she is not asking for the triggering cause, but for the structuring cause. The action of the technician is a structuring cause. It causes the conditions, in which the event of typing in a 50 causes the jukebox' playing 'Radio Gaga'.

Take another example: Edgar is waiting at a red traffic light. The moment the traffic light turns green he walks across the street. The question about the cause of his action, i.e., his crossing the street, can have two answers. The first answer is that Edgar's seeing the green light caused him to walk across the street. The second answer is that Edgar walks across the street the moment he sees the green light because he wants to safely cross the street. The first answer answers the question why Edgar walks across the street at this moment. It refers to the triggering cause of his action, i.e., his seeing the green light. The second answer answers the question why he walks across the street when he sees a green light, i.e., why his seeing the green light caused him to walk across the street. After all, he could have walked while the light was still red. Or he could have reacted differently to seeing the green light, for example, by closing his

eyes. The second answer gives a structuring cause, i.e., that Edgar wants to safely cross the street. It explains why seeing a green light causes Edgar to walk across the street.

Let me give one last example. This one is described by Fred Dretske:

> "A bell rings and a classically conditioned dog behaves the way it was conditioned to behave: it salivates. […] The Bell rings (*S*), and this produces a certain auditory experience (*C*) in the dog. The dog *hears* the bell ring. These sensory events, *as a result of conditioning*, cause saliva to be secreted (*M*) in the dog's mouth. What then, causes the dog to salivate? Well, in one sense, the ringing bell causes the dog to salivate. At least the bell, by causing the dog to have a certain auditory experience, triggers a process that results in saliva's being secreted into the dog's mouth. Yes, but that doesn't tell us why the dog is doing what it is doing – only why it is doing it *now*. What we want to know is why the dog is salivating. Why isn't it, say, jumping? Other (differently trained) dogs jump when they hear the bell. Some (not trained at all) don't do much of anything. So what causes the dog to salivate? This, clearly, is a request, not for the triggering cause of the dog's behavior, but for the structuring cause. It is the request for the cause of one thing's causing another, the cause of the auditory experience causing salivary glands to secrete."[186]

Dretske points out that the learning history of the dog explains why hearing the ringing of the bell causes the dog to salivate. Thus, the structuring cause of the dog's behavior is his having been trained like that.

How can we systematically explicate the concepts of triggering and structuring causes? As a first step, it is helpful to distinguish between the event that causes another event and the background conditions or enabling conditions that makes this kind of causal process possible. Seeing a green light causes Edgar to walk across the street only if he is not paralyzed, for example. Typing in a 50 causes the jukebox to play Radio Gaga only if the amplifier is not broken. One event causally triggers another event only if certain background or enabling conditions hold.

[186] Fred Dretske (1988): *Explaining Behavior. Reasons in a World of Causes* (Cambridge, MA: MIT Press), 43 f.

Are structuring causes something like background conditions? No, they are not. The fact that Edgar is not paralyzed is a background condition that makes it possible that his seeing the green light causes him to walk across the street. But it is not a structuring cause because it is not true that seeing the green light causes Edgar to cross the street *because he is not paralyzed.* Thus, the concept of background conditions or enabling conditions by itself does not help in explicating the concept of a structuring cause.

Let us get back to Dretske's understanding of a structuring cause. In the above quote, he states that a structuring cause is "the cause of one thing's causing another." A structuring cause makes it the case that the occurrence of a type-A event will cause a type-B event. It should be noted that the causal connection between type-A and type-B events is contingent on the structuring cause. Thus, in the absence of the structuring cause, type-A events are not causing type-B events.

Here is Dretske's more detailed explication (I substituted Dretske's abbreviations C and M, using instead the terms A-event and B-event):

> "In looking for the cause of a process, we are sometimes looking for the triggering events: what caused the [A-event] *which* caused the [B-event]. At other times we are looking for the event or events that *shaped* or *structured* the process: what caused [the A-event] *to* cause [the B-event] rather than something else. The first type of cause, the triggering cause, causes the process to occur *now*. The second type of cause, the structuring cause, is responsible for its being *this process*, one having [the B-event] as its product, that occurs now."[187]

Thus, triggering causes cause a certain event A, which in turn causes an event B; structuring causes cause the conditions in which A-type events cause B-type events.

My proposal is that intentions should be regarded as structuring causes of actions. Intentions structure the perception-action system[188] of an agent in such a way that it produces particular actions given a certain context. Which actions an agent performs depend on how she conceives of her situation. Intentions are a mental state

[187] Fred Dretske (1988), 42.
[188] Important for initiating action is how an agent conceives of her situation.

that determines which actions an agent performs given a certain situation. In other words, intentions cause the agent to perform a certain action if she conceives of her situation in a certain way. Edgar, for example, has the intention to walk safely across the street. This intention causes the conditions in which Edgar's seeing the green light causes his walking across the street.

The alternative picture views intentions as triggering causes of actions. According to this approach, the intention is like the ringing of the bell that causes the dog to salivate. If intentions were triggering causes, they would proximally cause the behavior, for example, a certain movement of the legs. Given this understanding, how would one describe the Edgar example? Edgar sees the green light. Seeing the green light causes an intention to walk across the street now. This intention, in turn, causes Edgar to walk across the street. According to this suggestion, sensory input indirectly causes an action by causing intentions.

Here are three considerations that speak in favor of the structuring cause account. The first one has to do with the relation between automatic behavior and control. Traditionally, automaticity and control have been thought of as opposites.[189] One consequence of such a view is that the range of behavior that lies outside of the agent's control becomes rather wide. On the basis of an alleged opposition between automaticity and agential control, people have formulated a strong challenge to the idea that we are in control over what we do. A widely received example is Daniel Wegner's "The Illusion of Conscious Will."[190] One major thread that runs through this work is the denial of agential control on the basis of a whole range of evidence that shows that most or maybe even all of our behavior is automatic.

This challenge rests firmly on the premise that automaticity and control are mutually exclusive. And this premise becomes plausible against the background of a triggering cause account of intentions. The thought is as follows: Actions are exercises of control. Agential control is realized by the proper functioning of intentions. In order to realize agential control, intentions have to cause the behavior.

[189] For example R.M. Shiffrin/W. Schneider (1977): 'Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory', in: *Psychological Review* (84), 127-190.
[190] Daniel M. Wegner (2002). Compare for a recent, critical psychological discussion of Wegner's arguments Roy F. Baumeister/E. J. Masicampo/Kathleen D. Vohs (2011): 'Do Conscious Thoughts Cause Behavior?', in: *Annual Review of Psychology* (62), 331-361.

Automatic behavior is not triggered by intentions. Hence, if intentions are triggering causes, automatic behavior is not under agential control.

Conceiving of intentions as structuring causes allows us to reject this line of thought. According to a structuring cause account, automatic behavior can be under agential control. What is required for this is that the causation of the behavior in this situation depends on an intention of the agent. Thus, compared to a triggering cause interpretation, understanding intentions as structuring causes leads to an ascription of more agential control. I regard this as an advantage.

Take, for example, Tim who drives home in his car. All of a sudden, a dog jumps into the street. Tim is reacting automatically: he hits the brakes immediately. In a situation like this, it is implausible to assume that seeing the dog causes an intention to hit the brakes, which in turn causes the behavior. If someone reacts automatically to a perceived stimulus, it is much more natural to regard this behavior as being directly triggered by the perception and not as being mediated by the sudden acquisition of an intention. According to the triggering cause account, then, Tim has no control over his action, i.e., hitting the brakes. But if intentions are structuring causes, then hitting the brakes could be under agential control if Tim has the intention to avoid running over dogs with his car and hence hits the brakes automatically as soon as a dog jumps in front of his car. And this latter interpretation seems to be quite plausible.

Second, the structuring cause account better fits to the empirical evidence. The psychologist Thomas Goschke suggests, based on a number of different studies about action control, "that we conceive of intentions as *constraints* that 'set the stage' for later processing, by configuring processing systems, increasing the sensitivity of processing pathways, and modulating the readiness of action schemas to be activated by subsequent stimuli."[191] In the terminology of the structuring cause account, this just is structuring the perception-action system in such a way that the agent reacts to certain situations with a particular kind of behavior. Goschke directly addresses the issue of automaticity and control:

[191] Thomas Goschke (2003): 'Voluntary action and cognitive control from a cognitive neuroscience perspective', in: Sabine Maasen/Wolfgang Prinz/Gerhard Roth (eds.) (2003): *Voluntary Action. Brains, minds, and sociality* (Oxford: Oxford University Press), 49-85, 58.

> "Within this framework, processing can be considered *automatic* to the degree that an intention can be realized by strong pre-existing connections between perceptual patterns and response programmes that were established by prior practice. Processing can be considered *controlled*, when (a) the selection of an intended response requires active maintenance of a representation of the current task or intention, because prepotent, but inadequate responses must be overridden, or when (b) completely new stimulus-response bindings must be established."[192]

(b) describes the central idea of the structuring cause account, namely, that an agent decides to react in a specific way to a particular kind of situation. Once this contingent "stimulus-response binding" is established, the behavior can be triggered automatically when the situation occurs. This latter idea is extensively discussed in Peter Gollwitzer's work on "implementation intentions," which I will discuss later in more detail.[193]

More evidence for the structuring cause account comes from brain imagining studies. Earl K. Miller and Jonathan D. Cohen investigate the role of the Prefrontal Cortex (PFC) in action control: "the PFC is important when 'top-down' processing is needed; that is, when behavior must be guided by internal states or intentions."[194] They understand guidance of behavior as the central function of intentions. And the way they understand the implementation of this guidance matches the structuring cause account:

> "we argue that all these functions [of different kinds of guidance] depend on the representation of goals and rules in the form of patterns of activity in the PFC, which configure processing in other parts of the brain in accordance with current task demands. These top-down signals favor weak (but task-relevant) stimulus-

[192] Thomas Goschke (2003), 61.
[193] Peter M. Gollwitzer (1993); Peter M. Gollwitzer/Caterina Gawrilow/Gabriele Oettingen (2010): 'The Power of Planning: Self-Control by Effective Goal-striving, in: R.R. Hassin/K.N. Ochsener/Y. Trope (eds.) (2010): *Self Control in Society, Mind, and Brain* (Oxford: Oxford University Press), 3-26.
[194] Earl K. Miller/Jonathan D. Cohen (2001): 'An Integrative Theory of Prefrontal Cortex Function', in: *Annual Review of Neuroscience* (24), 167-202, 168.

response mappings when they are in competition with more habitual, stronger ones."[195]

According to this picture, intentions guide action by "configuring processing in other parts of the brain," that is, by structuring the perception-action system in a certain way. Against this background, it would be false to conceive of intentions as triggers of action.

A third consideration that favors the structuring cause account over the triggering cause account is directly concerned with autonomy. One dimension of autonomy that many philosophers emphasize is that the autonomous agent defines, to a certain degree at least, what kind of person she is. "Many philosophers have thought that human autonomy includes, or perhaps even consists in, a capacity for self-constitution – a capacity, that is, to define or invent or create oneself."[196] Velleman mentions, among others, Charles Taylor[197], Christine M. Korsgaard[198], Harry Frankfurt[199], and Michael Bratman[200]. The central idea is that the autonomous agent is able to determine her own character or practical identity, that is, she is able to determine what kind of person she is. The structural cause account of intentions allows us to explicate a notion of partial self-constitution because it provides us with an idea of hierarchical control. By forming intentions, a person can exert control over her own character, that is, she can partially determine what kind of person she is. For example, if someone adopts the goal of becoming a good teacher, she can control her subsequent intention formation and execution in the light of this commitment, thereby reaching her goal and making it true that she is a good teacher. The triggering cause account is inferior in explicating how persons can determine their own practical identity or character because it lacks the resources to explicate how agents shape their own character by forming intentions regarding how they want to react to different kinds of situations.

---

[195] Earl K. Miller/Jonathan D. Cohen (2001), 170.
[196] J. David Velleman (2006): 'The Self as a Narrator', in: J. David Velleman (2006): *Self to Self. Selected Essays* (Cambridge: Cambridge University Press), 203-223, 203.
[197] Charles Taylor (1989): *The Sources of the Self: The Making of the Modern Identity* (Cambridge: Cambridge University Press); Charles Taylor (1985): *Human Agency and Language. Selected Papers 1* (Cambridge: Cambridge University Press).
[198] Christine M. Korsgaard (1996).
[199] Harry G. Frankfurt (1988).
[200] Michael R. Bratman (2007): *Structures of Agency. Essays* (Oxford: Oxford University Press).

4.6 Conclusion

In this chapter, I have presented a naturalistic account of agency and action in which the notion of rational control is central. I explicated how a causal theory of action can account for rational control, thereby refuting the challenge of missing control. That is, I showed that the idea of control does not fall outside the scope of an event-causal ontology. The motivation for this discussion was, of course, that the notion of natural autonomy presupposes a naturalistic account of agency and action. Natural autonomy conceives of autonomy as a natural property of persons. Autonomous agents are able to use their agency in special ways. I used the shaping-metaphor to illustrate this point. The autonomous agent is able to shape her life according to her own desires, beliefs, and values – against opposition. I already said that this shaping takes form through the agent's actions. Hence, this chapter provided us with an essential foundation for understanding the concept of autonomy I am interested in.

The central claim of the causal theory of action is that behavior is an action if it is caused in the right way by the agent's intentions. This causation constitutes rational control. An intention to x represents x-ing as an action that is either instrumental to achieving some goal G or is itself a goal of the agent. The structuring cause account of intentions says that intentions configure the perception-action system so that the agent performs particular kinds of action that are conducive for goal achievement when she perceives certain situational cues. The formation of intentions enhances the agent's control over her behavior because it allows her to make plans for the future, to plan with alternative scenarios, and to make good use of limited resources.

I emphasized that one of the core intuitions about autonomous agency is that the autonomous agent forms her own authentic standpoint and expresses it in her life. We made progress towards a better understanding of how one expresses one's standpoint. Against the background of the causal theory of action that I developed, we can say that expressing one's own standpoint in one's life is mediated by forming and executing intentions in accordance with this standpoint. But how can we make sense of the standpoint metaphor? This question has dominated the autonomy debate since Harry Frankfurt. In the next chapter, I discuss in more detail the most important attempts to answer this question.

5. Self-Directed Agency

Natural autonomy is autonomy within nature. It is a natural property of natural agents like us and not some non-natural add on that transcends our natural make up. In the last chapter, I presented an account of agency and action that explicates agency and action within a naturalistic ontology. This is an important contribution to an understanding of natural autonomy because it spells out the foundation of autonomy and the source of its value for us, namely our ability to shape our lives through our actions. But as important as it is, a naturalistic account of agency and action does not by itself tell us what autonomy consists in. The autonomous agent is not just able to act, she can make a special use of her ability to act. Autonomy refers to dispositions and abilities that constitute a more advanced kind of agency. What characterizes this kind of agency?

I have already pointed out that the autonomous agent is the agent who is able to shape her life in accordance with her own desires, beliefs, and values – and that she does this against opposition. In Chapter 2, I said that autonomy is concerned with expressing one's authentic standpoint. The autonomous agent is true to herself instead of following someone else's lead or acting on the basis of desires, beliefs, and values that are in some emphatic sense not really her own. Now this introduces something puzzling: isn't it the case that all of the agent's desires, beliefs, and values are hers? What kind of sense can we make of the idea that there might be a difference between what an agent wants and what she *really* wants? Questions such as these concern the issue of self-directedness.

In this chapter, I will systematically introduce the idea of self-directedness and discuss its importance for my understanding of autonomy. As I have pointed out shaping one's life consists in performing actions. The autonomous agent shapes her life in a special way. She expresses *her own* standpoint. The notion of self-directedness refers to action that has its source in the agent's own standpoint, and an account of self-directedness explains what this special kind of ownership consists in. In other words, an account of self-directedness explains what it means to have an authentic standpoint that grounds one's autonomous agency.

In 5.1, I discuss the rationale for a notion of self-directedness when explicating the concept of autonomy under consideration. I then proceed by discussing the most important approaches towards an understanding of self-directedness. As will become

apparent, I favor an account of self-directedness that puts the idea of an agent's practical identity in the center. Towards the end of this chapter, I will highlight some considerations that point beyond self-directed agency to resolute agency, the topic of Chapters 6 and 7.

5.1 Self-Directed Agency and Autonomy

Autonomy is closely tied to self-directedness. It appears to be a necessary condition for autonomous agency that the agent is self-directed. Marina Oshana formulates this idea succinctly when she writes: "Generally speaking, an autonomous person is one who is self-directed."[201] What exactly self-directedness consists in is a difficult question. It is arguably the central question in the autonomy debate for half a century now. Let us take a step back and ask why we need to introduce the notion of self-directedness in the first place. What are the phenomena that push us into the direction of distinguishing between self-directed agency and agency that is not self-directed? Let us think again about compulsive agency, like compulsive gambling or severe drug addiction. A compulsive gambler or a severe drug addict acts on desires from which she is, in some intuitive sense, alienated from. Here is another example: a mother who loves her child and is committed to a violence-free upbringing at some point loses control and hits her child. If she really despises violence, has always lived up to her ideal of a violence-free upbringing, and deeply regrets her action, it seems apt to say that she was not really herself in letting go of herself like this. Of course she is responsible for what she did. Nonetheless, there is a certain sense in which she was not properly self-directed. She did something that she didn't really want to do. This has a paradoxical ring to it because, after all, she did it, and if she did it, she was motivated to do it. And since 'wanting' is an umbrella term for all motivational states, it is apt to say that she wanted it. What could it mean, then, that she wasn't self-directed?

The pressure for developing a notion of self-directedness comes from such cases as compulsive and conformist agency. In these cases, the agent acts intentionally, but somehow she does not stand behind her action. It is intuitively appealing to view an agent who acts compulsively or conformist as being alienated from this action. This intuition is vague and in serious need of clarification. This is

[201] Marina A. L. Oshana (1998), 81.

precisely the function of an account of self-directedness. It tries to clarify these issues and to systematically account for our intuitions. Let us examine the connection between self-directedness and autonomy. The idea of self-directedness is essential for autonomy since autonomy gains its significance in the first place based on the assumption that there is a meaningful way to define an agent's standpoint. Only an agent who has her own standpoint can be autonomous because autonomy consists in being true to one's own standpoint. A lack of autonomy consists in either failing to express one's own standpoint or failing to develop one's own standpoint in the first place. The standpoint metaphor is pervasive in the autonomy debate. It illustrates the necessary starting point for any reasonable understanding of autonomy, namely, the possibility of a conflict between the agent and forces external to the agent. In order to make sense of the idea of autonomy, we first need to make sense of the idea that an agent has, or can develop, her own standpoint and, second, that this standpoint might be violated. An account of self-directedness is supposed to describe in a systematical way how an agent's standpoint is constituted and how it can be violated.

Another way to point at the questions that an account of self-directedness is supposed to answer consists in conjuring up the idea of an agent's self. We assume that it makes sense to distinguish between different ways in which a person can be related to her mental states, character traits, or actions, and that these differences ground the intuition that not everything an agent does reflects what she really wants to do. In other words, we assume that a sort of inner conflict is possible in which the agent takes sides or is positioned on one side and not the other. A way to describe this is to say that autonomous agency reflects, expresses, or springs from the agent's self, whereas nonautonomous agency is marked by a determination of something that is external to the self. The inclination to frame this difference in terms of an agent's true or real self certainly derives from the fact that the term 'autonomy' has been taken to refer to a self that is governing itself. The self vs. non-self distinction seems to be at the heart of autonomy, not least because etymology suggests this. On this reading, an account of self-directedness explicates the differences between self and non-self.

I have already said that the contemporary debate about personal autonomy is largely couched in this framework. Much of this is due to Harry Frankfurt who introduced the notions of "internal" and "external" desires for distinguishing between

self and non-self.[202] An agent counts as autonomous if and only if the desire that motivates her action is an internal one. External desires might move an agent to action, but in these actions, the agent is not autonomous. Here are some other examples: Gary Watson discusses the same problem when he alludes to the difference between "free action and intentional action."[203] So is Michael Bratman when he distinguishes "agential direction" and "agential governance."[204] In the same vein, Sarah Buss says: "*Autonomous* action, it is true, requires something more than the minimal self-direction intrinsic to mere intentional action."[205] And Laura Ekstrom, to mention just one last philosopher, explicitly refers to an "agent's *true* or *most central* self,"[206] which is active in autonomous action, in contrast to merely intentional actions that are motivated by "attitudes [that] do not represent what I accept to be the case or what I really want to do or desire."[207] The examples could be easily multiplied. Ekstrom certainly is right when she says that "[i]t is precisely the difficult issue of settling which forces are external and which are internal to the agent himself (which are 'truly his own') that is at the center of the discussion between Frankfurt and Watson and many others since."[208]

A closer look reveals that self-directedness is primarily discussed in contrast to compulsive action broadly conceived. Frankfurt set the tone for the whole debate when he introduced the case of the unwilling addict as his paradigmatic example for a lack of autonomy. Others, like Watson and Bratman, have followed his lead by highlighting that agents are internally split – an assumption that allows us to make sense of the idea that agents are not always self-directed, even though they are driven by motives that in one way or other belong to them.[209] In the following sections, I discuss the most important approaches towards an understanding of self-directedness. These approaches also exemplify what I call self-directedness accounts of autonomy

---

[202] For example Harry G. Frankfurt (1988 b): 'Identification and Externality', in: Harry G. Frankfurt (1988), 58-68.
[203] Gary Watson (1975): 'Free Agency', in: *Journal of Philosophy* (72), 205-220, 205.
[204] Michael R. Bratman (2007 b): 'Autonomy and Hierarchy', in: Michael R. Bratman (2007), 162-186, 177 f.
[205] Sarah Buss (1994), 95.
[206] Laura Waddell Ekstrom (1993): 'A Coherence Theory of Autonomy', in: *Philosophy and Phenomenological Research* (53), 599-616, 608.
[207] Laura Waddell Ekstrom (1993), 607.
[208] Laura Waddell Ekstrom (2005), 146. And since this discussion is predominantly about personal autonomy, the self vs. non-self distinction, in one guise or other, has been regarded by many as lying at the heart of an adequate understanding of personal autonomy.
[209] The two other paradigmatic cases of non-autonomy that I mentioned in Chapter 1, namely manipulation and coercion, have received only scant attention.

because they conceptualize autonomous agency as self-directed agency. Although the concept of autonomy that I am pursuing is richer than that because it also contains, in addition to the dimension of self-directed agency, the dimension of resolute agency, I agree that explicating the notion of self-directedness is an essential task in explicating an account of autonomy.

5.2 Harry Frankfurt – The Hierarchical Account

The historical overview in Chapter 2 ended with a brief discussion of Frankfurt's hierarchical account of autonomy. I emphasized that this is one of the central, if not *the* central contemporary approach towards personal autonomy. Frankfurt paved the way for an individualistic understanding of autonomy in which the idea of self-directedness takes center stage. At this point, I want to return to Frankfurt's position, present it in more detail, and discuss it critically. The central question that guides me is how self-directedness is spelled out in a Frankfurtian framework.

The core idea of Frankfurt is that an agent is autonomous with regard to her will when she endorses this will or identifies with it. Against the action theoretical background that I developed in Chapter 4, I would rephrase this as endorsing one's *intention*. This simple hierarchical account gives us a *prima facie* plausible explanation why, for example, the unwilling addict is non-autonomous. Her non-autonomy is due to the fact that she does not identify with her will. Quite to the contrary, she wants to have a different will. A similar explanation seems to apply to the aforementioned mother that hits her child and to other cases of non-autonomy. At first glance, the simple hierarchical account gives us a plausible account of self-directedness: an agent is self-directed if and only if she has the will she wants to have.

As it stands, however, this approach threatens to run into a regress. According to the hierarchical account, the will of an agent, that is, her motivating desire, is autonomous if and only if it is endorsed by a second-order volition. Now, this is compatible with a scenario in which an agent has a third-order volition to get rid of her second-order volition. I might have a second-order volition, for example, to act on my desire to hit the guy that insulted me. In addition, however, I might also have a third-order volition to get rid of this second-order volition, maybe because I view it as a destructive product of my troubled and violent upbringing. In this case, we would

have a conflict between a second-order and a third-order volition, and for this reason, it would be odd to say that the agent is self-directed if her will coheres with her second-order volition. After all, she wants to get rid of this second-order volition. Hence it appears that coherence between the first and the second order is not sufficient for grounding an agent's self-directedness. The second-order volition has to be endorsed by yet another higher-order volition. This threatens a regress of ever-higher steps in the hierarchy because there is, in principle, no end to the number of levels.[210]

We could answer this challenge by pointing out that, in fact, every hierarchy stops somewhere, and that it is sufficient for self-directedness that the will coheres with the higher-order volitions the agent *de facto* has. A refined proposal would then refer to higher-order volitions instead of second-order volitions and add an exclusion requirement: An agent is self-directed if and only if (i) she has a higher-order volition, V, that her motivating desire should be her will, and (ii) there is no conflicting higher-order volition, non-V, to get rid of V. This proposal stops the regress.

However, at this point a further problem arises. In virtue of what do the highest-order volitions possess any authority to speak for the agent? Let us call this the problem of authority. According to Frankfurt, the agent's standpoint is realized by her conflict-free highest order-volitions. But why should the highest-order volitions possess the authority to constitute the agent's standpoint? This authority cannot be inherited from yet another even higher higher-order volition since, by definition, these are the volitions of the highest order. Without further explanation, it remains mysterious why these unendorsed highest-order volitions should possess the special authority to determine an agent's autonomous standpoint. After all, in the case of conflict, we could also imagine that at least sometimes the agent has to be identified with the first-order desire instead of the higher-order volition. Indeed, this is a suggestion that Bernard Berofsky, for example, makes. "A higher-level desire need not be as deep or significant to the person as a first-order desire against which it is

[210] James Stacey Taylor, for example, formulates this regress problem that arises because the agent's autonomy with regard to her second-order desires is unclear. "If [an agent] is autonomous with respect to this second-order desire because it is, in turn, endorsed by a yet higher-order desire, then a regress threatens, for the question will then arise as to whether she is autonomous with respect to this third-order desire – and so on." James Stacey Taylor (ed.) (2005)), Introduction, 6.

directed."[211] This remark shows that just assuming that higher-order volitions define an agent's standpoint would beg the question

That it is not self-evident that the higher-order volitions constitute what the agent really wants is a thought that Gary Watson famously pointed out. He remarked that second-order volitions possess no inherent authority because, after all, they are just other desires of the agent. "Since second-order volitions are themselves simply desires, to add them to the context of conflict is just to increase the number of contenders; it is not to give a special place to any of those in contention."[212] The reason for this is that it remains unclear how a higher-order desire can ground an agent's self-directedness if a first-order desire by itself is not enough to ground it. As we have seen, one cannot simply point to yet another higher-order volition because this would start the regress again. The problem of authority is, in other words, to explain how a desire that is not endorsed by another, higher-order desire receives its authority to ground self-directedness. This question poses a dilemma for the simple hierarchical account. The first horn of the dilemma is that an answer to the problem of authority by appeal to a hierarchy of desires leads to a regress. The second horn of the dilemma is that any other answer shows the hierarchical account in and of itself to be incomplete, thereby questioning that higher-order volitions are necessary grounds for self-directedness in the first place. If authority at some level is not derived from reflective endorsement, why should it be derived at any level from it?[213]

The problem of authority lies at the heart of how to account for self-directedness. Not every intention that causes an action has the authority to speak for the agent. Hence, we need to give some criterion to distinguish those mental states that constitute the agent's standpoint from those that don't. Frankfurt comes forward with the idea that the notion of endorsement by a second-order volition provides us

---

[211] Bernard Berofsky (1995): *Liberation from the Self. A Theory of Personal Autonomy* (Cambridge: Cambridge University Press), 99.

[212] Gary Watson (1975), 218. Compare also: "higher-order volitions are just, after all, desires, and nothing about their level gives them any special authority with respect to externality. If they have that authority they are *given* it by something else. To have significance the hierarchy must be grounded in something else that precludes externality." Gary Watson (1987): 'Free Action and Free Will', in: *Mind* (96), 145-172, 149.

[213] Compare Watson's remark: "There may be something to the notions of acts of identification and of decisive commitment, but these are in any case different notions from that of a second (or n-) order desire. And if these are the crucial notions, it is unclear why these acts of identification cannot be themselves of the first order – that is, identification with or commitment to courses of action (rather than with or to desires) – in which case, no ascent is necessary, and the notion of higher-order volitions becomes superfluous or at least secondary." Gary Watson (1975), 219.

with such a criterion. But without further argument, this appears to be an arbitrary stipulation. Indeed, Frankfurt concedes that "the assignment of desires to different hierarchical levels does not by itself provide an explanation of what it is for someone to be *identified* with one of his own desires rather than another."[214] In other words, the problem of authority remains unsolved. As a consequence, Frankfurt faces the challenge of how to augment the simple hierarchical account of self-directedness in a non-arbitrary way so that he can explain why some hierarchies ground self-directedness. He answers this challenge by introducing the notions of identification, wholeheartedness, and satisfaction. According to the broadened account, an agent is self-directed if and only if she is moved by a desire that she wholeheartedly identifies with.[215] The idea is, then, that wholeheartedness conveys authority. What does it mean to identify with a desire wholeheartedly? And does this augmentation avoid the regress and really solve the problem of authority?

In order to explicate the notions of identification, wholeheartedness, and satisfaction, Frankfurt refers to a special kind of inner conflict. According to Frankfurt, there are two basic kinds of inner conflict, but only one of them pertains to questions of identification and wholeheartedness. The first kind of inner conflict is the conflict between first-order desires and higher-order volitions. If I desire to smoke a cigarette and also desire that this desire does not prevail in determining my actions, I have conflicting first-order and second-order desires. As troubling as it might be, this kind of inner conflict is irrelevant for questions about identification and wholeheartedness. For these matters, a second kind of inner conflict becomes prevalent, namely, the conflict between different second-order volitions, or any higher-order volitions for that matter. If I have a desire to get rid of my desire for donating a considerable amount of my monthly payment and also have a desire to be moved by my desire to donate, I have conflicting second-order volitions. This second kind of conflict constitutes the absence of wholeheartedness. An agent is wholeheartedly endorsing a first-order desire if and only if she has no conflicting higher-order desires regarding the first-order desire in question.[216]

[214] Harry G. Frankfurt (1988 c): 'Identification and Wholeheartedness', in: Harry G. Frankfurt (1988), 159-176, 166.
[215] Compare Harry G. Frankfurt (1988 c), 175.
[216] Compare Harry G. Frankfurt (1988 c), 165.

Frankfurt points out that "[w]holeheartedness does not require that a person be altogether untroubled by inner opposition to his will. It just requires that, with respect to any such conflict, he himself be fully resolved. This means that he must be resolutely on the side of one of the forces struggling within him and not on the side of any other."[217] Wholeheartedness, in turn, is explicated in terms of satisfaction. "In what does his wholeheartedness with respect to these psychic elements consist? It consists in his being fully satisfied that they, rather than others that inherently (i.e., non-contingently) conflict with them, should be among the causes and considerations that determine his cognitive, affective, attitudinal, and behavioral processes."[218] Satisfaction "does entail […] an absence of restlessness or resistance."[219] "Satisfaction is a state of the entire psychic system – a state constituted just by the absence of any tendency or inclination to alter its condition."[220] In other words, satisfaction consists in a harmonious set of higher-order volitions.

The following passage is a concise summary of Frankfurt's augmented account of autonomy as self-directedness:

> "On hierarchical accounts, a person identifies with one rather than with another of his own desires by virtue of wanting to be moved to action by the first desire rather than the second. […] The mere fact that it is a second-order desire surely gives it no particular authority. […] The endorsing higher-order desire must be, in addition, a desire with which the person is *satisfied*. […] Identification is constituted neatly by an endorsing higher-order desire with which the person is satisfied."[221]

Given the aforementioned understanding of satisfaction, according to which an agent is satisfied with a higher-order desire if and only if she lacks any tendencies or inclinations to get rid of it, we can say that an agent is self-directed if and only if (i) she is moved by a desire that she desires to be her will, (ii) she has no conflicting

[217] Harry G. Frankfurt (1999 a): 'The Faintest Passion', in: Harry G. Frankfurt (1999), 95-107, 100.
[218] Harry G. Frankfurt (1999 a), 101.
[219] Harry G. Frankfurt (1999 a),101.
[220] Harry G. Frankfurt (1999 a), 104.
[221] Harry G. Frankfurt (1999 a), 105.

higher-order volitions, and (iii) she has no tendencies or inclinations to alter her higher-order volitions.

The augmented hierarchical account avoids the problem of regress because it does not require that every element in the hierarchy be endorsed on an even higher level. According to the augmented account, there is no need for ever higher layers in the hierarchy because, if an agent lacks conflicting higher-order volitions concerning a particular first-order desire, and if she has no tendencies or inclinations to alter her higher-order volitions, she is satisfied with her higher-order volition, which in turn constitutes her wholehearted endorsement of her first-order desire.

However, this reply is still vulnerable to the problem of authority. What is Frankfurt's answer to that problem? The augmented hierarchical account points to the direction of an answer by emphasizing that an agent who is wholehearted lacks any fragmentation of the self. The self is "unified,"[222] and thus it contains no desire that could undermine self-directedness. At this point, Frankfurt merges his hierarchical account of desires with a specific notion of a person's self or identity. According to this notion, a person shapes her own practical identity by wholeheartedly identifying with a first-order desire. "The decision [to terminate a sequence of desires or preferences of increasingly higher orders] determines what the person really wants by making the desire on which he decides fully his own. To this extent the person, in making a decision by which he identifies with a desire, *constitutes himself*."[223] And if such a desire, which is constitutive of the person, moves her to action, she is self-directed because she is directed by that which constitutes her self.

Of course, this account of self-creation and practical identity is far too sketchy as it stands. One wants to know, for example, how this decision can be made: does the agent need certain capacities? What are the starting points for the self-constitution? Another line of questioning concerns the limits or boundaries of this process: can I give myself any identity I want to have? If not, are there any pre-existing building blocks of a person's identity which are unaffected by my decisions? And yet another set of problems concerns questions such as: why should a desire not be part of my identity only because I want to get rid of it? Isn't it a truism that everybody needs to accept that part of her identity are unwanted desires? However, putting aside these

[222] Harry G. Frankfurt (1999 a), 100.
[223] Harry G. Frankfurt (1988 c), 170.

general and deep issues concerning questions about practical identity, it becomes clear that Frankfurt possesses, at least in principle, an answer to the problem of authority. The answer is, in short, that wholeheartedly endorsed desires constitute the agent's self and thus ground her self-directedness. The volitional structure wholeheartedness consists in is not just arbitrarily picked. Its significance, in Frankfurt's understanding, lies in the fact that it constitutes the agent's practical identity.

Now, although I have just highlighted problems and open questions concerning Frankfurt's approach, I also want to emphasize that Frankfurt gives us, in outline at least, a plausible explanation why compulsive action renders the agent non-autonomous. His answer is that the compulsive agent acts out of motives that do not belong to her self. With this answer, Frankfurt brings the idea to the foreground that the autonomous agent is able to shape her life on the basis of desires, beliefs, and values that are in some emphatic sense expressive of her. Because of this, the hierarchical account is a valuable source for exploring and explicating the concept of autonomy I am interested in. After all, as I pointed out in the beginning, compulsion is indeed a threat to autonomy and, hence, an account of autonomy should be able to explain why this is so. As I also explained there are other paradigmatic cases of non-autonomy, like coercion and manipulation. Frankfurt places the problem of compulsion in the center of the self vs. non-self distinction. How helpful is Frankfurt's hierarchical account in illuminating other paradigmatic cases of non-autonomy?

Being coerced to do something is one of the paradigmatic examples for non-autonomous agency. Someone else imposes her will onto you. This surely violates your autonomy. At first glance, the hierarchical account offers an explanation for the fact that autonomous agency is thwarted by coercion. Imagine Robbed Robbie who is the victim of a street robber. The robber threatens to kill him if Robbed Robbie does not hand over his purse. In giving away his purse, Robbed Robbie is non-autonomous. Why? Following the hierarchical account, one might say that Robbed Robbie's motivating desire, i.e., the desire to give away his purse, is not endorsed by a higher-order volition. After all, Robbed Robbie prefers to keep his money. This inner tension makes him non-autonomous, one might argue. On closer inspection, this explanation proves to be too hasty. It is plausible to assume that Robbed Robbie forms a second-order volition to comply with the threat. After all, compliance is his best strategy to

survive. It is perfectly rational for him to give away his money if this improves his chances of survival. Of course, it could be the case that Robbed Robbie desires not to be moved by his desire to give away his money, even though this seems to be quite irrational. However, the point is that even if Robbed Robbie is rational and desires wholeheartedly to be moved by his desire to hand over his money, he still acts non-autonomously in giving away his money. The hierarchical account lacks the resources to explain this. Moreover, according to the hierarchical account, Robbed Robbie autonomously hands over his money if he thinks that this is the smartest move to make in his circumstances and accordingly desires to be moved by a correspondent desire. But this is puzzling: how can it be the case that Robbed Robbie acts autonomously when he gives in to the threat? This would be an altogether different concept of autonomy.

The hierarchical account also has problems with the case of manipulation. According to the hierarchical account, self-directedness is just a matter of the structure of an agent's desires at a certain point in time. However, it appears to be possible that this structure is the very consequence of manipulation. But it surely is implausible that an agent who is thoroughly brainwashed such that she endorses her first-order desires wholeheartedly is autonomous. At least, this is the case for the concept of autonomy I am interested in. According to it, a perfectly brainwashed person is a paradigmatic case of a person that has lost her autonomy. Imagine Cultist Kurt who was thoroughly brainwashed by his cult with the result that he wholeheartedly desires what the cult leader expects of him. Cultist Kurt is wholehearted. There is no ambivalence in his higher-order volitions. However, due to the fact that this harmony in his psychology is the consequence of severe manipulation, he is a paradigmatic example of an agent lacking in autonomy. The hierarchical account cannot explain why Cultist Kurt is non-autonomous. Quite to the contrary, it would count Cultist Kurt as being autonomous because he wholeheartedly endorses his will. Again, an account of autonomy that views actions as autonomous that are the product of manipulation has a very different concept of autonomy in mind. This problem cannot be avoided by simply adding another element into the psychic

structure because no matter how complex the psychic structure becomes, it can always be imagined to be the product of a manipulation.[224]

What this shows us is that the hierarchical account is insufficiently equipped to illuminate why coercion and manipulation are instances of non-autonomy according to the concept of autonomy under consideration. It seems that we either need to modify Frankfurt's idea of what an agent's authentic self consists in or add an additional element in our account of autonomy if we want to adequately deal with these cases of non-autonomy. As will become apparent in Chapters 6 and 7, I will argue that we should not restrict ourselves to an understanding of self-directedness if we want to explicate autonomy. Instead, we should embrace a broader account that acknowledges the role of resoluteness in autonomous agency.

As an account of autonomy, the hierarchical account is incomplete. It is not easily transferable from cases of compulsion to cases of coercion and manipulation. In the next section I argue, following Gary Watson, that it is also incomplete as an account of self-directedness because it neglects the role of the agent's values in determining her standpoint. Another potential problem concerns Frankfurt's focus on local autonomy. As I defined in the first chapter, local autonomy is autonomy with respect to particular judgments, choices, actions and so forth. I pointed out that an account of autonomy should be able to account for both local as well as dispositional autonomy and also be able to spell out their interrelatedness. It seems that the hierarchical account is not well suited to being extended to an account of dispositional autonomy. The reason for this is that Frankfurt is not concerned with the role of the agent in bringing about wholeheartedness. The agent does not need to play an active role in creating a volitional structure that counts as wholehearted. Hence, we cannot identify a set of abilities and dispositions that make her an autonomous agent. The dispositionally autonomous agent is disposed to be active in a certain way. This element of activity is lost in emphasizing a passive property like wholeheartedness.

To sum up: Frankfurt develops a hierarchical account of self-directedness according to which an agent is self-directed if and only if her will is reflectively

[224] Gary Watson rightly points out that this problem is pervasive: "the other, more conspicuous problem [the problem of manipulation] we noted with hierarchical accounts does seem to reveal an inherent difficulty for the entire spectrum of compatibalist theories. Again, this is the problem of origins of one's will, however that is to be understood. These theories say nothing to rule out the possibility that one's evaluations, or higher-order volitions (or brute acts of identification) are merely the products of conditioning, manipulation, or brainwashing." Gary Watson (1987), 151.

endorsed by a higher-order volition with which the agent is satisfied. Satisfaction is understood as the absence of any tendency or inclination to alter one's higher-order volitions concerning the will in question. In short: if the agent endorses her will without conflicting higher-order volitions, she is wholehearted and thus self-directed. The lack of self-directedness is either due to the fact that the agent's motivating desire is in conflict with the agent's higher-order volitions or to the fact that the agent lacks harmonious higher-order volitions in the first place. The augmented hierarchical account of self-directedness avoids the problem of a regress by denying that every desire has to be necessarily affirmed on a higher level in order to ground self-directedness. The problem of authority can be solved if one grants Frankfurt's claim that identification and wholeheartedness constitute the very self that is at stake when it comes to questions of self-directedness. The cases of coercion and manipulation, however, cannot be sufficiently explained by the hierarchical account. Although these open questions and problems remain, Frankfurt's concepts of identification, wholeheartedness, satisfaction, and endorsement have shaped the discussion about personal autonomy and self-directedness significantly.

The discussion shows that the hierarchical account captures something very important about our volitional structure. We are able to take a stance towards our own desires and motives, and it is quite common that we in fact take such a stance. For Frankfurt, such a stance is solely a matter of the agent's desires. In the next section, I present Gary Watson's critique of this claim. Following Watson, I argue that the agent's standpoint is constituted by her values.

5.3 Gary Watson – The Evaluational Account

In contrast to Frankfurt, Gary Watson highlights the importance of the agent's values and normative reasons for self-directed agency. His approach views a normative perspective as a central condition for making sense of the idea that an agent might be internally divided in a self and a non-self, because against this background, we can imagine a conflict between my normative perspective and my desires. And in such a conflict, the agent's standpoint in the emphatic sense that is conjured up by talking about self-directedness is constituted by her normative reasons and her reason-based judgments – or so Watson claims. How is this supposed to work?

Gary Watson is, like Frankfurt, primarily concerned with how to make sense of the problem that "what one most wants"[225] is not always what one does.[226] He mentions as exemplary cases in which one does something that one does not want to do actions "which are explained by addiction, manias, and phobia."[227] These examples show that Watson, just as Frankfurt, works on a solution to the problem of self-directedness with the aim to explain how autonomous agency is distinct from compulsive action. He starts by sketching a "platonic conception of practical reasoning,"[228] which distinguishes between "Reason and Appetite,"[229] both of which are understood to be sources of motivation. Watson refers to this difference in terms of "*judging* good" or "valuing" on the one hand, and "desiring" on the other.[230] Once this distinction is in place, we can see what it can be to claim that what one really wants is different from what one desires most.

> "The answer is that the phrase 'what one most wants' may mean either 'the object of the strongest desire' or 'what one most *values*.' […] The problem of free action arises because what one desires may not be what one values, and what one most values may not be what one is finally moved to get."[231]

Once the distinction between value judgment and desire is in place, we have the resources to make the self vs. non-self distinction.

The agent's self or standpoint is determined by her value judgments, which can be in conflict with her desires. An important background assumption is that desiring to do something does not imply that one values it. The case of compulsion proves to be instructive here because, in severe forms of compulsion, the agent can have a strong desire to act in a certain way without valuing the desired action at all.

---

[225] Gary Watson (1975), 209.
[226] In his "Free Action and Free Will" Watson makes it explicit that from a systematic point of view, freedom contains two aspects, namely "self-determination (or autonomy) and the availability of alternative possibilities." Gary Watson (1987), 145. According to this understanding, autonomy is self-determination and his distinction between an evaluational standpoint and desires is supposed to give us an account of self-determination (or self-directedness as I call it here).
[227] Gary Watson (1975), 205.
[228] Gary Watson (1975), 207.
[229] Gary Watson (1975), 207.
[230] Gary Watson (1975), 208.
[231] Gary Watson (1975), 209.

"What is distinctive about such compulsive behavior, I would argue, is that the desires and emotions in question are more or less radically independent of the evaluational system of these agents."[232] Whereas desiring does not imply a corresponding value judgment, valuing implies a corresponding desire. That is, if an agent judges a certain course of action to be good or the best, she automatically develops the desire to act accordingly.[233] An agent's values are understood as follows: "We might say that an agent's values consist in those principles and ends which he – in a cool and non-self-deceptive moment – articulates as definitive of the good, fulfilling, and defensible life. […] we all have more or less long-term aims and normative principles that we are willing to defend. It is such things as these that are to be identified with our values."[234] This means that the agent's concept of a good life expresses her values. Building on this idea Watson introduces the notion of an agent's valuational system. "*The valuational system* of an agent is that set of considerations which, when combined with his factual beliefs (and probability estimates), yields judgments of the form: the thing for me to do in these circumstances, all things considered, is *a*."[235] In other words, an agent's values guide her judgments about what she ought to do.

Mirroring the platonic distinction between Reason and Appetite, there also exists a counterpart to the valuational system: the motivational system. "*The motivational system* of an agent is that set of considerations which move him to action. We identify his motivational system by identifying what motivates him."[236] Given that 'desire' has become an umbrella term for all mental states that are at least partly constituted by motivation, the motivational system is the set of all desires. As we have already observed, it is not necessary that every desire be accompanied by a corresponding judgment that finds the object of the desire to be good. In fact, however, both desire and value judgment often go hand in hand.[237]

---

[232] Gary Watson (1975), 220.
[233] In a later article, Watson doubts that the connection between judging good and desiring is as straightforward as he presents it here. He concedes that it appears indeed possible to judge something to be good without developing a corresponding desire. "Notoriously, judging good has no invariable connection with motivation, and one can fail to 'identify' with one's evaluational judgements. One can in an important sense fail to value what one judges valuable."
[234] Gary Watson (1975), 215.
[235] Gary Watson (1975), 215.
[236] Gary Watson (1975), 215.
[237] "Now, to be sure, since to value is also to want, one's valuational and motivational systems must to a large extent overlap." Gary Watson (1975), 215. Watson takes this overlap to be guaranteed because he assumes, at this point, that to judge valuable is sufficient for desiring – an assumption that he drops in his later work.

According to Watson's valuing account of self-directedness, an agent is self-directed if and only if she acts in accordance with her valuational system. Self-directedness consists in performing an action because one judges it to be the action that, all things considered, one ought to perform. Watsons argues in a way similar to Frankfurt when he claims that the special authority of the valuational system steams from its role in constituting an agent's practical identity.

> "One's evaluational system may be said to constitute one's standpoint, the point of view from which one judges the world. The important feature of one's evaluational system is that one cannot coherently dissociate oneself from it *in its entirety*. […] In short, one cannot dissociate oneself from all normative judgments without forfeiting all standpoints and therewith one's identity as an agent."[238]

Watson himself sheds light on this passage in a later article: "Only evaluations can give one reasons to oppose first-order desires, and when and only when agents' behaviour expresses their evaluations are they sources and 'authors' of (because they 'authorized') their behaviour."[239] The importance of an agent's evaluational standpoint lies in its authority to determine the agent's standpoint in cases of inner conflict. And it has this authority because it is constitutive of what the agent regards as good, right, and valuable. There is no sensible way to dissociate an agent's evaluational standpoint from her self-directedness. Indeed, it sounds odd to ask whether an agent who does what she judges to be good, right, and valuable and does it because she judges it to be good, right, and valuable is self-directed. Of course, one might insist that such a standpoint might be distorted or delusional and hence without authority. But if it is not in doubt that the agent's judgment is free of delusions or self-deception, there appears to be no room left for doubting that the agent is self-directed when acting accordingly.

How plausible is the claim that only an agent's values provide her with a standpoint in the emphatic sense needed for self-directedness? I think that this claim is

[238] Gary Watson (1975), 216.
[239] Gary Watson (1987), 149.

indeed highly plausible. Here is the argument as Watson develops it. He starts out with the premise that we need some criterion to distinguish between desires that express the agent's standpoint and those that don't. He then asks how we can account for the idea that an agent can stand in opposition to her own desires. He reckons that it cannot be something that is intrinsic to the desires themselves because from the perspective of the desire, each of them is equally valid. Hence, in order to introduce the right kind of conflict, we need values.

Watson's evaluational account of self-directedness is successful in providing a possible and plausible explanation for what distinguishes autonomous and compulsive agency. Can his account be extended to the case of coercion? How would one explain why Robbed Robbie lacks autonomy in giving away his purse when threatened by a street robber? A straightforward explanation runs like this: Robbed Robbie does not value giving away his money. This action is not backed up by his evaluational system. Hence, when he does it, he acts against his evaluational standpoint. And acting against one's evaluational standpoint is non-autonomous agency. However, as attractive as this explanation might sound in the first place, it runs into similar problems as we have encountered when trying to apply Frankfurt's hierarchical account to the case of coercion. Of course it is possible that the action of giving away the money is not supported by Robbed Robbie's evaluative judgment. But although possible, it is highly implausible since it is rational for Robbie not to gamble with his life. We can assume that he values living and that he is interested in continuing to live. At the same time, we can assume that the loss of his purse, though inconvenient, does not have any catastrophic consequences. I think that it is plausible to assume that, in general, people believe that it is a good idea to give in to a threat when your life is at stake and all you lose by complying with the threat is some money. But even if you reason in this manner and judge that you ought to give away your purse, your action is rendered non-autonomous by the fact that you are doing it because you are threatened. An account of autonomy should be able explain why this is the case. And Watson's evaluational account fails to meet this requirement.

How does Watson handle the case of manipulation? In his "Free Action and Free Will," Watson concedes that his evaluational account lacks the resources to deal with the problem of manipulation. He envisions what he calls "Brave New World

cases."[240] The common feature of Brave New World cases is that the intrinsic conditions of autonomy, whichever these may be according to your preferred theory, are brought about by "conditioning, manipulation, or brainwashing."[241] Watson does not give a fully-fledged treatment of this problem. But at least he hints at a possible solution by giving an analysis of why people are impaired in their autonomy in Brave New World cases: "The crucial thing about their situation is that they are incapable of effectively envisaging or seeing the significance of certain alternatives, of reflecting on themselves and on the origins of their motivations, of comprehending or responding to relevant theoretical and evaluational criteria."[242] And since "[i]t is part of our idea of autonomy that the fundamental determinants of our behaviour are ones that we could endorse without delusion,"[243] every satisfactory account of self-directedness needs to address the question of under which conditions an agent is in the relevant sense free of delusions. Watson himself refrains from saying anything more about this.

Another issue that lacks clarity concerns cases in which the agent's evaluational standpoint appears to be non-autonomous. Conformist action, for example, can express a conformist evaluational standpoint. But the conformist agent is not the autonomous agent. Watson could answer that self-directedness always consists in expressing one's evaluational standpoint, whatever that may be. Even a conformist evaluational standpoint grounds self-directedness. Autonomy, however, needs more then self-directedness. Or he could try to argue that the agent is not self-directed in cases such as conformism and that it is a mistake to construe an agent's evaluational standpoint in this way.

I will leave this question open for the moment. The central insight of Watson is that being true to one's evaluational standpoint is a necessary condition for self-directedness and that the all-things-considered judgment of an agent is the major determinant of her standpoint as an autonomous agent – maybe even the sole determinant. I follow Watson in assuming that autonomy becomes an issue if and only if an agent is able to develop her own standpoint based on values and normative reasons. Only an agent who possesses values in addition to mere desires has the

[240] Gary Watson (1987), 151.
[241] Gary Watson (1987), 151.
[242] Gary Watson (1987), 152.
[243] Gary Watson (1987), 153.

internal complexity that is necessary for the kinds of conflicts autonomy is concerned with. The central idea in this context is that an agent cannot be alienated from her normative reasons in the way she can be alienated from her desires because we cannot make sense of the idea that an agent who is free of undue influence and who forms an intention to x on the basis of her judgment that she ought to x does not stand behind her will. Desires, in contrast, can always be opposed by an agent's values and normative reasons.

In summary, it can be stated that Watson's approach offers us a convincing way to spell out the self vs. non-self distinction. Moreover, his evaluational account allows us to understand why compulsive agency undermines autonomy. The reason is that the compulsive agent violates her evaluational standpoint. However, as before when we discussed Frankfurt, we need to concede that Watson does not shed light on the reasons why coercion and manipulation undermine autonomy. In the following chapters, I argue that this shortcoming is based on the neglect of resolute agency. But before I open that debate, let me continue with discussing the role of self-directedness in autonomy. It is natural to introduce Michael Bratman at this point since his own account of self-directedness and autonomous agency is an advancement of the frameworks of Frankfurt and Watson.

5.4 Michael Bratman – The Planning Account

Both Frankfurt and Watson use the notion of an agent's self or identity as the foundation on which to build an account of self-directedness. However, their remarks about how an agent's self is constituted remain rather superficial. One philosopher who tries to close this gap is Michael Bratman, who develops his own planning account of self-directedness and autonomy in close contact with Frankfurt's and Watson's approaches.

We are planning agents. This insight lies at the heart of Michael Bratman's philosophical endeavor.[244] His theory of planning agency has become standard fare in action theory. Recently he has further developed this planning approach towards agency into a theory of autonomous agency.[245] This latter work is significantly influenced by the works of Harry Frankfurt and Gary Watson. He uses their

[244] Michael R. Bratman (1987).
[245] Michael R. Bratman (2007).

frameworks and theoretical insights as starting points and develops them further. Just like Frankfurt and Watson, Bratman is systematically firmly placed in the business of giving an account of the self vs. non-self distinction. He inherits the problem of how to distinguish between what an agent does and what she really wants to do. As a consequence, the distinction between merely intentional action and autonomous agency moves into the focus of his work.

Here are some terminological clarifications that we need in order to understand Bratman correctly. Bratman introduces the notion of "agential direction." Agential direction is the consequence of what I call rational control. Bratman explicates it as follows: "As a first step we can say that for the agent to direct thinking and acting is for relevant attitudes that guide and control that thinking and action to have authority to speak for the agent – to have agential authority. […] When relevant attitudes with such agential control appropriately guide and control, the agent directs."[246] The authority Bratman has in mind is not the authority of the agent's autonomous standpoint, but of the agent's intentional standpoint. Intentional agents have an intentional perspective. Bratman's concept of agential direction accounts for intentionality in a naturalistic way without introducing some sort of homunculus. With this concept Bratman identifies non-agential parts of the agent that constitute agency. In this sense, they possess the authority to speak for the agent. This kind of authority, however, grounds all intentional action, no matter whether it is autonomous or not. Agential direction consists in guidance by mental states or processes that constitute the agent's perspective as such. But agential direction is not sufficient for self-directedness.

Agential direction is a necessary condition for autonomous agency. But autonomy is more demanding. Bratman distinguishes "self-governance" from mere agential direction and points out that autonomy, as he understands it, is captured in terms of self-governance.[247] The crucial difference between agential direction and self-governance is that "in self-governance the agent herself directs and governs her

[246] Michael R. Bratman (2007 a): 'Introduction', in: Michael R. Bratman (2007), 3-18, 4. It is important to notice that Bratman wants to remain in a broadly naturalistic framework and that he aims, for this reason, at identifying mental states, structures, or processes whose functioning constitutes an agent's engagement in action.

[247] "When I talk of autonomy it is, in particular, this idea of self-governance that is my direct concern." Michael R. Bratman (2007 a), 4.

practical thought and action."[248] Self-governance requires, in addition to intentionality, that the effective motive have "subjective normative authority." For autonomous agency, we need guidance by normative reasons.

> "For the agent to *govern* her thinking and acting, however, it is not sufficient that she directs them. To govern is to direct in a way that is shaped by what the agent treats as justifying considerations, as reasons. In self-governance, attitudes that have agential authority need to guide relevant thought and action by way of articulating what has, for the agent, justifying significance – what has normative authority for the agent."[249]

In other words, an agent governs herself if and only if she performs an action because she believes that she has best reason to perform this particular action. Self-governance implies, whereas agential direction does not, that the agent acts as she does because she thinks that there is some consideration that speaks in favor of doing so. The agent takes herself to be justified in performing this action.

> "Agential governance is a particular form of […] agential direction: agential governance is agential direction that appropriately involves the agent's treatment of certain considerations as justifying reasons for action. Autonomous action involves a form of agential direction that also constitutes agential governance."[250]

By emphasizing the importance of normative reasons, Bratman takes sides with Watson against a Frankfurtian picture, according to which an agent's autonomy is only a matter of the structure of her desires.

---

[248] Michael R. Bratman (2007 a), 4.
[249] Michael R. Bratman (2007 a), 4 f.
[250] Michael R. Bratman (2007 b), 177. Compare also: "There is agential direction of action when action is under the control of attitudes whose role in the agent's psychology gives them authority to speak for the agent, to establish the agent's point of view – gives them, in other words, agential authority. This agential direction of action is, furthermore, a form of agential governance of action only when these attitudes control action by way of the agent's treatment of relevant considerations as justifying reasons for action, that is, as having subjective normative authority for her." Michael R. Bratman (2007 b), 177 f.

In order to explicate in more detail what self-directedness consists in, Bratman introduces the notion of "self-governing policies."[251] According to Bratman, "policies are intentions that are appropriately general in their content. They support treating, over time, like cases in like ways, and doing this is a matter of (and so, with reference to one's) policy."[252] A policy says how to react to or how to evaluate certain states of affairs. For example: if I have to drive and I find myself having a desire for drinking a glass of wine, I don't act on this desire. Of special importance for self-governance are those policies that directly concern practical deliberation, that is, policies which determine which considerations ought to be treated as speaking in favor for or against different types of action.

> "According to the intention-based theory, then, a kind of valuing that is at the heart of self-governance consists in policies (that is, general intentions) of giving weight or other forms of significance to certain considerations in practical reasoning and action. Call these *self-governing policies*."[253]

A self-governing policy could consist in, for example, regarding one's desire for chocolate cake as providing me with a reason to order it when possible. Self-governing policies contain both aspects that are needed for self-governance: agential direction and subjective normative authority. "Policies that say what to treat as a reason, and with what weight and significance –and thereby help determine what has subjective normative authority – can bring together, in the way needed for self-governance, both agential and subjective normative authority."[254] Hence, an agent is self-governed, or self-directed, if her actions are guided by her self-governing policies. What is so special about self-governing policies? Why should self-governing policies possess the authority to ground an agent's self-governance? Bratman uses the argumentative structure that we already encountered in Frankfurt and Watson. He argues that self-governing policies constitute the agent's self or standpoint. Two considerations back up this claim.

---

[251] Michael R. Bratman (2007 a), 6; Michael R. Bratman (2007 d): 'Three Theories of Self-Governance', in: Michael R. Bratman (2007), 222-253, 239.
[252] Michael R. Bratman (2007 a), 6.
[253] Michael R. Bratman (2007 d), 239.
[254] Michael R. Bratman (2007 a), 6.

First, self-governing policies are important constituents of an agent's personal identity. They secure continuity and connection between elements of the agent's psychological household over time. For this reason, it is justified to attribute their workings directly to the agent. Bratman explicates this idea within a Lockean framework of personal identity.

> "Indeed on a broadly Lockean approach to personal identity, the connections and continuities that are the back-bone of this psychological, cross-temporal quilt are constitutive of the identity of the agent over time, an identity that is presupposed in much of our practical thinking. And this suggests the conjecture that it is primarily its role in constituting and supporting this organized, cross-temporal, Lockean interweave of action and practical thinking that confers on a structure of attitudes a claim to speak for the agent – a claim to agential authority."[255]

Self-governing policies constitute personal identity. Hence, behavior that is guided by self-governing policies is guided by the agent.[256] Second, self-governing policies have subjective normative authority because their content specifies what to treat as a practical reason. It also assigns significance to those reasons. To use a Watsonian notion: self-governing policies determine the agent's "evaluational standpoint."

One might question the claim that subjective normative authority is necessary for self-governance. As we have seen, Frankfurt's account of self-directedness rejects the claim that something like subjective normative authority or an evaluational standpoint is necessary for self-directedness. For Frankfurt, a psychological structure that is more like Bratman's agential authority is sufficient for self-directedness. Watson has argued against Frankfurt by pointing out that the very idea of a difference between the strongest motivation and what an agent really wants requires reference to an evaluational standpoint. In the same vein, Bratman tries to motivate his more demanding notion of self-governance. He points out, "the very idea of governance brings with it, I think, the idea of direction by appeal to considerations treated as in

[255] Michael R. Bratman (2007 a), 5.
[256] In a later article, Bratman adds the condition that the agent has to be satisfied with this policy: "To have agential authority, we can say, a self-governing policy must be one with which the agent is, in an appropriate sense, satisfied." Michael R. Bratman (2007 c): 'Planning Agency, Autonomous Agency', in: Michael R. Bratman (2007), 195-221, 210.

some way legitimizing or justifying. This contrasts with a kind of agential direction or determination that does not involve normative content."[257] Frankfurt's hierarchical account is concerned with a type of the latter kind of agential direction. Hence, at this point, Bratman follows Watson in his Frankfurt critique.

In brief summary, Bratman regards self-governing policies as grounding self-directedness. A self-governing policy specifies what considerations to treat as a reason and with what significance to refer to them in one's practical deliberation. Given a broadly Lockean picture, self-governing policies are constitutive of an agent's personal identity because they support the psychological interconnectedness that constitutes an agent's identity. And for this reason it is correct to view them as representing the agent. Hence, they underlie "agential direction." In addition, self-governing policies possess subjective normative authority because they determine what the agent regards as normative. Taken together, these two properties guarantee that self-governing policies are at the heart of self-governance – or self-directedness for that matter.

How does this theory explain the problematic cases? Compulsive actions, like those of the unwilling addict, are motivated by a desire that the agent does not regard as reason giving. Addicted Andy does not have a self-governing policy to treat his desire for drugs as reason giving. If this desire becomes effective, it lacks subjective normative authority. The lack of subjective normative authority does not imply that the agent, Addicted Andy for example, does not act intentionally. Intentional action requires what Bratman calls agential direction. And surely, in this minimal sense, compulsive action can be directed by the agent. However, the agent lacks self-directedness in the emphatic sense. According to Bratman's account, Addicted Andy is not self-governed in taking drugs, and this is exactly the result we expect. Compulsive action is non-autonomous, and Bratman provides us with a possible explanation why this is so.

The case of coercion proves to be more difficult. Robbed Robbie is non-autonomous when he gives in to the threat and hands over his money. What is the desire that drives him to action? A good candidate for this desire is the desire to remain alive or to stay unharmed. We can safely assume that this desire possesses

[257] Michael R. Bratman (2007 c), 209.

subjective normative authority for Robbed Robbie. Every mentally healthy human adult has a self-governing policy to treat her desire for staying alive and remaining physically unharmed as reason giving. It might not be the decisive reason in every situation, but it certainly is always a relevant consideration. Given that Robbed Robbie has this kind of self-governing policy, and given that his effective desire is the desire to stay alive, then he appears to be self-governing. His effective desire possesses subjective normative authority. This result, however, is unsatisfying because we agreed at the outset that enforced actions are non-autonomous. We want an explanation why this is so. Bratman's account falls short of this demand. To be sure, we could amend the case of Robbed Robbie in a way that makes it accessible for Bratman. Let us imagine that Robbed Robbie is in the grip of a very strong fear. When he gives away his money, he is driven by terror. Now in this case, it is less plausible to assume that his action is backed up by a self-governing policy to treat one's terror as reason giving. And if Robbie lacks such a policy, his effective desire lacks subjective normative authority. Again, this is the result that we wanted. Hence, the amended case strengthens Bratman's account. However, even granted that Bratman has an explanation for the amended case, this does not nullify the problems with the original case. It certainly is a possible scenario that someone gives in to a threat and that her effective desire is treated as reason giving by the agent. The amended case is a variation that does not render the original case as being irrelevant.

What this shows is that Bratman shares the same problem that we encountered with Frankfurt and Watson. His approach allows that agents who act under coercion are locally autonomous. But this contradicts the concept of autonomy under consideration as I explicated at the very beginning of this discussion. The same problem occurs with respect to the case of manipulation. Bratman admits quite frankly that he lacks an explanation for the case of manipulation. A manipulated agent might be indistinguishable from a non-manipulated agent at a certain moment in time. To distinguish them, we need to consider their histories. Bratman assumes that a comprehensive account of autonomy needs to spell out historical requirements for autonomous agency. However, this is not what he aims at. He describes his project as getting a clear account of the structural aspects of autonomous agency. And he argues that we need an understanding of these necessary conditions of autonomous agency before we can sensibly investigate what kind of additional historical requirements for autonomous agency exist.

To sum up: Although Bratman's conceptualization of an agent's self or her authentic standpoint is more advanced than those of Frankfurt and Watson, he still falls short of illuminating the problematic cases of coercion and manipulation. However, his notion of self-governing policies is an important contribution to an understanding of the intuition that the autonomous agent shapes her life according to desires, beliefs, and values that are in some emphatic sense expressive of herself. A self-governing policy is partly constitutive of the agent's self. Hence, we can explicate part of the concept of autonomy under consideration by saying that the autonomous agent acts in accordance with her self-governing policies.

5.5 Laura Waddell Ekstrom – The Coherence Account

I have just emphasized the value of Michael Bratman's account of what constitutes an agent's identity. An alternative conception of the self, which was also developed with an eye to autonomy, is that of Laura Waddell Ekstrom. Ekstrom develops a "Coherence Theory of Autonomy." She calls it "a real-self approach to autonomy"[258] because it aims at accounting for autonomy in terms of governance by one's "real-self." Accordingly, an agent counts as autonomous if and only if she is moved by her real self. Ekstrom argues that this real self is constituted by a set of cohering mental states. For this reason, she dubs her theory a *coherence* theory of autonomy. The coherence theory is an account of self-directedness. Ekstrom makes this explicit when she writes: "We all agree that autonomy is to be understood as self-direction, self-command, or self-rule."[259] She points out that "'self-direction' means direction by the genuine or true self – and not the inauthentic or contrived or externally imposed self."[260] The self vs. non-self distinction lies at the heart of Ekstrom's understanding of autonomous agency.

Ekstrom starts with the idea that "[t]o be autonomous is to be self-governed. An individual acts autonomously when he acts on his own reasons. Thus, in order to understand autonomous action, we must know what it is for a reason to count as *one's own*; and thus we need a conception of what constitutes 'the self'."[261] We can distinguish between two claims here. First, autonomous agency is agency that is

[258] Laura Waddell Ekstrom (2005), 152.
[259] Laura Waddell Ekstrom (2005), 155.
[260] Laura Waddell Ekstrom (2005), 155.
[261] Laura Waddell Ekstrom (1993), 599.

guided by one's own reasons. Second, if we want to understand what one's own reasons are, we need a conception of one's real self. Ekstrom does not argue extensively for these claims. The first claim is shared by many philosophers. We already encountered it in the approaches of Watson and Bratman. The second claim is more controversial. Still, Ekstrom moves directly forward to take up the challenge that the second claim implies, namely to explicate what an agent's true self consists in. She tries to illuminate this notion of a self by distinguishing it from those 'parts' of an agent which, though attributable to the agent in a "weak" sense, are not truly her own.

> "We must acknowledge that all states of me, mental and physical, are *mine*, so in some sense, they are all part of 'the self.' But I am concerned with exploring a moral notion of the 'self,' rather than giving a metaphysical account of personal identity. My interest is in knowing what the core of a moral agent is, what constitutes one's central or *true self*."[262]

In a later article this emphasis on issues of morality is somewhat weakened. There she also uses the notion of a psychological self in order to describe what she is interested in: "the idea is that certain of our attitudes are more central to who we are in a moral or psychological sense than are other of our attitudes, and that it is in acting on these more central attitudes that we exert special direction over our lives."[263] This psychological or moral sense underpins her talk of a real, true, or most central self. We should note that Ekstrom rejects any attempts to hypostatize a self. "None of the proponents of such an [real-self] approach, as I understand them, including myself, intends to propose an ontological thesis, according to which, for instance, there is some sort of entity – some item to be added to our metaphysics – floating around or somehow attached to the human being."[264] The real self consists in a set of mental states. It is not an entity over and above these mental states.

What mental states constitute the real self? In order to answer this question, Ekstrom introduces the notion of a "preference." She defines a preference "as a desire

[262] Laura Waddell Ekstrom (1993), 600, FN 5.
[263] Laura Waddell Ekstrom (2005), 153.
[264] Laura Waddell Ekstrom (2005), 152 f.

that has survived a process of critical evaluation – in particular, with respect to an individual's conception of the good."[265] That is, preferences are desires that the agent deems to be good or valuable. She gives the following definition:

> "A *preference*, as I shall use the term, is a very particular (or peculiar?) sort of desire: it is one (i) for a certain first-level desire to be effective in action, when or if one acts, and (ii) that is formed in the search for what is good. A preference, that is, has as its intentional object the state of affairs of a certain of one's first-level desires being satisfied in action, and is formed by an agent's *evaluating* that first-level desire with respect to some standard of goodness."[266]

Taken literally, this definition implies that every preference is formed by a deliberative process. This, however, appears to be too strong. After all, we certainly acquired many of our desires through non-deliberative processes. If none of these desires could be part of the true self, we would exclude a lot of plausible candidates for a membership in the true self. I think that it is more reasonable to broaden the second requirement and allow that desires that withstand a deliberative process, although this process was not part of their etiology, can be preferences.

Ekstrom sees her concept of a preference as an enriched version of Frankfurt's notion of a second-order volition. Preferences are, in contrast to Frankfurt's second-order volitions, necessarily based on a process of evaluation, that is, they are retained with respect to an agent's considerations about her reasons and about her conception of the good.

Preferences are important because they are partly constitutive of an agent's real self. To understand this claim, we need to contemplate Ekstrom's notion of a character and a character system. "I propose that we take any given self to be a particular character together with the power for fashioning and refashioning that character, where the character or what I call the character system, of an agent S at a time t is the set of the propositions that S accepts at t and the preferences of S at t."[267]

[265] Laura Waddell Ekstrom (2005), 148.
[266] Laura Waddell Ekstrom (1993), 603.
[267] Laura Waddell Ekstrom (1993), 606.

Hence, a self is constituted by acceptance states (a notion that Ekstrom borrows from Keith Lehrer), preferences, and some basic faculty of self-formation. This self is not the real self. The real self is a subset of these preferences and acceptances, namely those that cohere with each other:

> "So far I have defended the claim that an agent's self should be taken to be, together with her evaluative faculty, not all of her desires and beliefs, but rather a subset of these: those that she acquires and retains in her attempt to believe what is true and to desire what is good; that is, her acceptances and preferences. Now I wish to make the proposal that we take an agent's *true* or *most central self* to be a subset of these acceptances and preferences, namely, those that *cohere* together. One's preferences, I suggest, are *authorized* – or sanctioned as one's own – when they cohere with one's other preferences and acceptances."[268]

That a preference is authorized means that it has the authority to speak for the agent. Authorized preferences determine an agent's evaluational standpoint. An agent is self-directed if she acts on the basis of an authorized preference.[269] Ekstrom gives an explicit definition of what she means by authorization: "df, personal authorization: *S* is personally authorized at *t* in preferring that *d* be effective in action if and only if the preference for *d* coheres with the character system of *S* at *t*."[270] Against this background, she formulates her "coherence account of autonomous action" as follows:

> "An act is autonomous just in case it is nondeviantly caused by an uncoercively formed, personally authorized preference. A preference that is personally authorized for an individual has authority for speaking for her, for representing what she truly wants, in being well supported by a network of her considered

[268] Laura Waddell Ekstrom (1993), 608.
[269] "When I act on an authorized preference, I act in a way that is autonomous." Laura Waddell Ekstrom (1993), 614.
[270] Laura Waddell Ekstrom (1993), 612. Compare also the alternative formulation: "df, personal authorization: *S* is personally authorized at *t* in preferring that *d* be effective in action if and only if every preference that competes with the preference for *d* for *S* on the basis of the character system of *S* at *t* is defeated or neutralized on the basis of the character system of *S* at *t*." Laura Waddell Ekstrom (1993), 612.

> attitudes; it is an attitude with respect to which she is wholehearted. Thus, action on such a preference is self-directed or self-ruled, rather than heteronomous."[271]

Ekstrom's understanding of self-directedness is a demanding one. The real self is just a fraction of the self as a whole. And in contrast to Watson's evaluational account, for example, the coherence account does not only rule out desires as autonomous which are not backed up by a value judgment, but also desires that the agent judges to be good but which are in tension which her coherent set of preferences. Hence, Ekstrom excludes any action as a possible candidate for autonomous action with respect to which the agent is ambivalent. There is no room for ambivalence in the real self. For this reason, it appears that the coherence account of autonomy is about a different concept of autonomy. After all, according to the concept of autonomy under investigation, we can distinguish different levels of autonomy and we can view actions as being autonomous which are not expressive of someone's real self. The coherence account lacks the resources to explain these judgments. It appears to describe a kind of ideal agency, but fails to explain differences in autonomy in non-ideal agents.

Apparently Ekstrom has a different concept of autonomy in mind. For her, autonomy is marked by some sort of ideal unity of mental states, in particular, of the agent's preferences. This idea does not help us to understand better the concept of autonomy that I investigate. To repeat, I am interested in autonomy as a gradual property that can be learned and improved, a kind of autonomy that is essentially concerned with situations of conflict and how an agent prevails in them. Before I tackle this idea directly, let me finish this chapter with a discussion of what I regard as the best way to account for the self vs. non-self distinction. The next section deals with the idea of an agent's practical identity as it is developed by Korsgaard and Charles Taylor.

5.6 Practical Identity. Christine Korsgaard and Charles Taylor

There are two central insights of the foregoing discussion. First, self-directedness is a matter of developing and expressing one's authentic evaluative standpoint. Second, in

[271] Laura Waddell Ekstrom (2005), 151 f.

order to account for the idea of an agent's authentic evaluative standpoint, we need to explicate a notion of the agent's self or identity as a person. The latter task is only superficially dealt with in Frankfurt's and Watson's work. Bratman and Ekstrom address it more directly. For Bratman, the role of self-governing policies is essential, and Ekstrom argues that a coherent set of preferences constitutes the core of the agent's self. I have already pointed out that I think that Ekstrom's coherence account conceptualizes the self too narrowly. In addition, her approach does not allow us to conceive of self-directedness as a gradual property.

Bratman's approach is more promising but fails, as I see it, in integrating the basic idea of why self-governing policies are essential for an agent's authentic standpoint into a broader picture. In order to make progress in this direction, I will conclude my investigation of the current debate with a discussion of Korsgaard's idea of practical identity. I will also mention Charles Taylor's contribution to such an understanding.

As I see it, the most promising approach towards understanding the idea of an authentic standpoint and its connection to autonomous agency can be extracted from Korsgaard's account of moral autonomy. As I discussed in Chapter 2, Korsgaard argues that normativity has its source in autonomy. Autonomy, in turn, is understood in a Kantian spirit as a form of self-legislation. The notion of the agent's practical identity is central in Korsgaard's framework. In my understanding, we can use the basic idea of an agent's practical identity in order to explicate what an agent's authentic standpoint consists in. But we need to get rid of the moral implications that Korsgaard connects with this idea.

An agent's practical identity is a picture of the kind of person one aspires to be. Korsgaard describes it as the "description under which you value yourself."[272] The practical identity is a conception of oneself as a particular kind of person. It grounds the agent's assessment of actions, projects, and ways to life. The roles the agent commits herself to are an important part of an agent's practical identity. If the agent commits herself to being a good parent, for example, this commitment shapes her deliberations about what she ought to do. As a good parent, she needs to spend time with her children and care for their wellbeing. Committing oneself to being a teacher,

[272] Christine M. Korsgaard (1996), 101.

a priest, a doctor, a politician, a lover, and so forth, brings with it a framework of practical reasons. Korsgaard captures this point in the following remark about autonomy and practical identity: "Autonomy is commanding yourself to do what you think would be a good idea to do, but that in turn depends on who you think you are."[273]

I pointed out above that Korsgaard views "moral identity" as a necessary part of an agent's practical identity. That is, the autonomous person cannot avoid understanding herself as a moral agent, thereby committing herself to morality. I think that this Kantian heritage gives us a misleading picture of personal autonomy as we nowadays intuitively understand this notion. Antigone, for example, or Marie Curie, are not conceived of as autonomous because of their commitment to morality. To be sure, they are committed to certain values and norms, and their autonomy lies in the fact that they live up to these values and norms even though this led them into serious trouble. But it is not necessarily a commitment to morality that is at stake here. Moreover, there exists no conceptual necessity to assume that a commitment to morality is implied by their commitment to other values and norms. We therefore need to relax Korsgaard's morality requirement for autonomy. In contrast to Korsgaard, who wants to develop a framework of normativity and its sources, we don't need to put any special emphasis on morality. I can fully agree with Korsgaard's claim that at the heart of autonomy lies self-government, and that self-government, in turn, depends on your practical identity. I just want to use this concept in a broader way that does not limit it to moral self-conceptions.

The notion of a practical identity allows us to account for an important aspect of autonomy, namely that the autonomous agent binds herself towards certain courses of action. Watson only points to the agent's evaluational standpoint without saying much about its genesis or the structure in which it is embedded. The notion of a practical identity elaborates on these issues. It explicates how an agent binds herself to certain values and norms by making them into parts of her practical identity. This account shares the focus on self-directedness but widens its scope by also including thoughts about the context in which self-directedness emerges and gains its importance for autonomous agency.

[273] Christine M. Korsgaard (1996), 107.

We can find a similar approach towards practical identity in Charles Taylor's work. In his paper "What is human agency?", he sets out to investigate "the notion of a self,"[274] building on Frankfurt's hierarchical framework. He agrees with Frankfurt that it is a distinctive feature of persons that they possess the ability to evaluate their own motives. But in contrast to Frankfurt, he thinks that we need to distinguish "between two kinds of evaluation of desire."[275] Taylor criticizes Frankfurt for being concerned only with quantitative evaluation, that is, evaluation based on the strength of a desire. For persons, however, a different mode of evaluation is much more essential, namely evaluation on the basis of qualitative aspects. In the center of Taylor's own approach stands this distinction between two modes of valuing, namely "weak evaluation" and "strong evaluation."[276] Taylor clarifies that weak evaluations are not quantitative in the sense that they express all alternatives "in some common units of calculation."[277] "All these weak evaluations are only 'quantitative' in the weak sense that they do not involve qualitative distinctions of worth."[278] A more precise way to phrase the difference, then, is by their relation to the idea of worth.

Taylor suggests two criteria for making the distinction between weak and strong evaluation:

> "(1) In weak evaluation, for something to be judged good it is sufficient that it be desired, whereas in strong evaluation there is also a use of 'good' or some other evaluative term for which being desired is not sufficient; indeed some desires or desired consummations can be judged as bad, base, ignoble, trivial, superficial, unworthy, and so on.
>
> It follows from this that (2) when in weak evaluation one desired alternative is set aside, it is only on grounds of its contingent incompatibility with a more desired alternative. […] But with strong evaluation that is not the case. Some desired consummation may be eschewed not because it is incompatible with another, or if because of incompatibility this will not be contingent."[279]

[274] Charles Taylor (1985 a): 'What is human agency?', in: Charles Taylor (1985), 15-44, 15.
[275] Charles Taylor (1985 a), 16.
[276] Charles Taylor (1985 a), 16.
[277] Charles Taylor (1985 a), 17.
[278] Charles Taylor (1985 a), 17.
[279] Charles Taylor (1985 a), 18.

We already encountered the first criterion when we discussed the difference between a brute desire and valuing. We said that it is possible to desire something without thinking that it is good or valuable. Taylor uses a different terminology to express the same idea. For weak evaluation, it is sufficient to have a desire. Strong evaluation, in contrast, is concerned with conceiving of something as valuable. The second criterion points to the fact that some desires or courses of action are perceived as being intrinsically good or bad. Based on a strong evaluation, we might decide to refrain from performing a particular action because of its intrinsic unworthiness. In weak evaluation, alternatives are only rejected because something else is more appealing, that is, object of a stronger desire.

The idea of strong evaluation is essentially tied to what I call, following Korsgaard, an agent's practical identity. Taylor gives us as an example for a strong evaluation based on which an agent rejects a particular action. He asks us to imagine an agent who rejects a particular action because she conceives of it as "a cowardly act." This rejection is based on the agent's practical identity. "If we examine my evaluative vision more closely, we shall see that I value courageous action as part of a mode of life; I aspire to be a certain kind of person. This would be compromised by my giving in to this craven impulse."[280] Strong evaluations are made against the background of the agent's practical identity. That is, the agent commits herself to living up to certain ideals and values as an intrinsic part of being "a certain kind of person."

Against this background, Taylor distinguishes two "kinds of self"[281] that correspond to the two kinds of evaluation, namely a "simple weigher of alternatives" and a "strong evaluator."[282] A central distinguishing feature is that the strong evaluator employs a "vocabulary of worth"[283] to articulate her evaluation of different desires or actions, whereas "the simple weigher's experiences of the superiority of A over B are inarticulable."[284] As a consequence of the employment of an evaluative vocabulary, the reflection also becomes "deeper." "A strong evaluator, by which we mean a subject who strongly evaluates desires, goes deeper, because he characterizes

[280] Charles Taylor (1985 a), 19.
[281] Charles Taylor (1985 a), 23.
[282] Charles Taylor (1985 a), 23.
[283] Charles Taylor (1985 a), 24.
[284] Charles Taylor (1985 a), 24.

his motivation at greater depth."[285] This metaphor is in need of clarification, and indeed Taylor gives us one that links strong evaluation with the agent's aspiration to be a certain person. "Motivations or desires do not only count in virtue of the attraction of the consummations but also in virtue of the kind of life and kind of subject that these desires properly belong to."[286] And he continues:

> "But this additional dimension can be said to add depth, because now we are reflecting about our desires in terms of the kind of being we are in having them or carrying them out. Whereas a reflection about what we feel like more, which is all a simple weigher can do in assessing motivations, keeps up as it were at the periphery; a reflection on the kinds of beings we are takes us to the centre of our existence as agents. Strong evaluation is not just a condition of articulacy about preferences, but also about the quality of life, the kinds of beings we are or want to be. It is in this sense deeper."[287]

Against this background, I understand self-directedness as being motivated in accordance with one's strong evaluations. And the perspective of a strong evaluator is determined by the agent's practical identity, that is, by her commitment to certain values and norms that are embedded in certain roles, characters, and kinds of life.

Does this suggestion run into a regress? After all, the commitment to a practical identity or aspects of a practical identity already reflects the values of the agent. Put like this, it seems that an agent's practical identity cannot be the *source* of her values, but only a particular *expression* of them. I think that it is misleading to conceive of the situation in terms of a dichotomy between a point in time at which the person has no practical identity and a subsequent point at which she has herself committed to a particular practical identity. In my understanding, a practical identity emerges in a rudimentary form once an agent starts to view herself as subject to certain normative demands ("A good girl brushes her teeth and says 'thank you'!") and it evolves over time. This evolution is a process that has reflexive aspects. The agent can think about her outlook on herself and change it. She reflects on parts of her

[285] Charles Taylor (1985 a), 25.
[286] Charles Taylor (1985 a), 25.
[287] Charles Taylor (1985 a), 26.

practical identity within an evaluational framework that is intrinsic to other parts of her practical identity. Part of this process is driven forward by a tension between certain desires and the agent's practical identity. In such a situation, the agent can either dismiss the desires as being an authentic part of her, or she can reconsider her practical identity and accommodate it. Another important motor of evolution is the acquisition of new values that are in tension with some older values, thereby necessitating a readjustment of one's practical identity. In this respect the evolution of an agent's practical identity is comparable to reaching a reflective equilibrium.

As I just stated I understand self-directedness as agency that is expressive of and has its source in the agent's practical identity. I think that this is compatible with Watson's idea of self-directedness as acting in accordance with one's evaluational standpoint. The notion of a practical identity helps us to define an agent's evaluational standpoint. In this respect, it deepens Watson's account. Bratman's framework of planning agency gives us some systematical building blocks of what a practical identity can consist in. In particular, the notion of self-governing policies captures something very important about practical identity. An essential part of the practical identity is a commitment to certain roles or attitudes. These roles or attitudes determine, at least partly, an evaluative standpoint, that is, they say what considerations to take as a reason and with what to consider them in one's deliberations. Bratman's notion of a self-governing policy seems to be slightly more restricted because Bratman claims that a self-governing policy says what desires to treat as reason giving, and this connection to the agent's desires is not necessary for all reason-giving considerations in the practical identity. But the basic idea of self-governing policies, namely, to commit oneself to a certain evaluational or normative framework, thereby generating starting points and constraints for one's deliberation, is identical to the reason-giving function of the practical identity.

Even the practical identity approach towards self-directedness falls short of explaining the whole range of cases of diminished autonomy. It is possible to conceive of an agent's practical identity as being the consequence of manipulation. And we can also coherently imagine that an agent's action of giving in to a threat is not contradicted by her practical identity. Resoluteness is not conceptually tied to practical identity. I will therefore explore and explicate the notion of resolute agency

separately in Chapters 6 and 7. Before I come to that, let me conclude this section by answering the challenge of the missing agent.

5.7 Answering the Challenge of the Missing Agent

In Chapter 3, I presented what I called the challenge of the missing agent. This challenge is directed at theories that try to account for agency and action within a naturalistic framework. The central idea of this challenge is that we cannot account for an agent's involvement in action when we only refer to the causal functioning of parts of the agent. To say, for example, that agency consists in the proper functioning of the agent's intentions leaves the agent out of the picture – at least according to this challenge. In this chapter, we have seen how the challenge of the missing agent can be refuted. What we need for this is an argument that shows us how it is possible that certain mental states of an agent possess the authority to represent the agent as a whole. We encountered many such arguments. For Frankfurt, an agent has to be identified with her internal desires. Watson identifies the agent with her evaluational standpoint, and Bratman with her self-governing policies. I followed Korsgaard in putting the idea of an agent's practical identity in the focus of any attempts to explain which mental states represent the agent.

Regardless of how one answers the question of which mental states represent the agent, we can refute the challenge of the missing agent as long as we believe that there is one correct answer. The basic idea is that an agent is constituted by a subset of her mental states. These make up the agent's self. Hence, when they are causing a particular action, it is the agent who causes it because she has to be identified with these mental states. Thus, my answer to the challenge of the missing agent is that the agent is fully active in performing an action if those mental states that cause the action belong to her practical identity. We can account for agency and action in naturalistic terms.

5.8 Conclusion

According to the concept of autonomy I explore in this study, the autonomous agent is characterized by expressing her own authentic standpoint. Indeed, it appears to be a central intuition about autonomy that the autonomous agent is the agent who stands

fully behind her actions. Her life reflects her fundamental desires, beliefs, and values. The starting point for an investigation of self-directedness is the intuition that it makes sense to distinguish between different ways in which an agent is related to her intentions. Some of her intentions are supposed to express in a deep way what kind of person she is. Others violate her innermost character or standpoint. Against this background, autonomy is conceived of as being essentially concerned with an expression of one's authentic standpoint. I have already pointed out that situations of compulsive, or manipulated, or coerced agency play a central role in backing up this intuition and the subsequent theoretical discussion.

I argued that an agent's authentic evaluative standpoint is determined by her practical identity. An agent's practical identity is that description of herself as a person under which she values herself (Korsgaard), or a description of the person she aspires to be (Taylor). It implies values and norms. And it is these values and norms that determine the agent's authentic standpoint. Self-directedness consists in a match between the demands of one's practical identity and the intentions and actions of the agent.

An account of self-directedness is an important part of an adequate explication of autonomy. But it needs to be supplemented for two reasons. First, self-directedness is a local matter. An explication of autonomy solely in terms of self-directedness only focuses on what makes an agent autonomous with respect to a particular action or behavior more broadly conceived. When it comes to the idea that autonomy is a dispositional property of persons, self-directedness accounts remain mute. There is no obvious route from the claim that being self-directed is an important aspect of local autonomy to an explication of dispositional autonomy. While self-directedness accounts of autonomy emphasize a very important aspect of local autonomy, they have problems when it comes to an understanding of dispositional autonomy. In other words, self-directedness accounts neglect to discuss that autonomy is a global matter, that is, that it refers to a dispositional property of persons.

A second and related problem concerns the fact that self-directedness accounts focus solely on the authentic standpoint of the agent when accounting for local autonomy. But this gives us too simple a picture of autonomy. In particular, this approach fails to give credit to the fact that autonomy is not only concerned with the expression of an authentic self or standpoint, but also with abilities and dispositions to

resist undue influences and to endure in conflicts. These two failures are connected with each other as will become apparent in the next chapters. Dispositional autonomy is constituted by dispositions and abilities to resist undue influences and to endure in conflicts. And local autonomy does not only consist in expressing one's authentic self or standpoint, but also in staying resolutely on one's path.

The value of self-directedness accounts is that they systematically investigate the intuition that autonomy is concerned with the deepest character of a person and its authentic expression. But there is more to autonomy than that as will become apparent in the next chapter. In it, I present some considerations that bring some deeper problems of self-directedness accounts to the surface while I at the same time broaden our understanding of the concept of autonomy under discussion.

6. The Limits of Self-Directedness

Self-directedness is an important dimension of autonomous agency. I argued for this claim in the last chapter and also discussed what self-directedness consists in. According to my understanding, self-directed agency is agency that has its source in the practical identity of the agent, that is, agency which is backed up by the agent's evaluational standpoint. But although an account of self-directedness is essential for explicating the concept of autonomy that I am interested in, there is more to autonomy than just self-directedness. I mentioned this several times throughout the discussion, in particular in Chapter 5, when we encountered some of the shortcomings of self-directedness accounts of autonomy. Against this background, I claim that we need to acknowledge the essential role of resolute agency in order to adequately explicate the concept of autonomy under investigation. In this chapter, I continue to explore aspects of autonomy that cannot be accounted for solely in terms of self-directedness. The common theme that runs through this chapter is that autonomy is concerned with how well an agent is equipped to deal with opposition and conflict. This antagonistic dimension of autonomy is insufficiently explained in self-directedness accounts. According to the concept of autonomy under discussion, autonomy is partly constituted by an agent's abilities and dispositions to overcome obstacles and to resist pressure. This is what constitutes dispositional autonomy. And this dimension of autonomy is not reducible to an account of self-directedness. My central claim, then, is that autonomous agency encompasses resolute as well as self-directed agency.

In Chapter 1, I presented a particular kind of conflict, namely coercion, as one paradigmatic case of non-autonomy. As became clearer in the last chapter, self-directedness accounts of autonomy are ill equipped to explain why coercion undermines autonomy. And this observation can be generalized: self-directedness accounts are insufficient when it comes to understanding those dimensions of autonomy that have to do with persisting in conflicts. This is the topic of this chapter. I present four considerations that expose the shortcomings of understanding autonomy solely in terms of self-directedness. All of these are concerned with autonomy as an aptitude to be an effective agent, even in the face of obstacles and opposition.

6.1 Analogy to Political Autonomy

In Chapter 1, I discussed the political origin of the notion of autonomy. I emphasized that the concept of autonomy under discussion has inherited a central feature of the original political understanding of autonomy, namely, that autonomy is concerned with conflicts, in particular with an agent's aptitude to prevail in different kinds of conflicts. And this is something that goes beyond mere self-directedness. For this reason, it is problematic to reduce autonomy to self-directedness. We get a first glimpse of this problem when we consider the analogy between personal and political autonomy.

As I discussed above, political autonomy, in its original understanding, has two dimensions. First, it has a normative reading and refers to the right of a *polis* to determine certain matters of its political and economic life for itself. In this sense, autonomy can be violated by interference in the *polis*'s internal affairs. This sense also grounds the demands some *poleis* raised to be free of certain kinds of interventions by external powers. It was this idea of autonomy as a right that Aigina invoked against Athens. When Sparta took sides with Aigina's position, the Peloponnesian Wars started.

The second dimension is concerned with the properties of a *polis* that constitute its aptitude in dealing with opposition and conflict. Thucydides informs us that certain military facilities were considered to be essential. An autonomous *polis* had its own fortifications and fleet. The fortifications were important for a *polis* to defend itself against attacks. The fleet also contributed to the defense. Additionally, it was a tool to secure the *polis*'s interests on a broader scale. One noteworthy aspect of this early idea of political autonomy is that autonomy described the status of a *polis* within a context of possible political and military conflicts. Autonomy describes the capacity to prevail in a possible threatening situation.

An interesting observation in this context is that political autonomy is not tied to democratic institutions. Democracy is not a necessary condition for autonomy. This conceptualization of political autonomy has preserved itself until today. Undemocratic states can be autonomous. We intuitively think of states like 18th-century monarchist France, Fascist Germany during the Second World War, or autocratic China as autonomous states. This intuitive understanding also conforms to international laws. Undemocratic states are typically treated as autonomous states by international law

and also in foreign affairs. Let me illustrate this with a particular example: pre-revolution Libya. Gaddafi's Libya was a dictatorship. The people lacked a number of democratic rights and freedoms. They were not allowed to vote freely or to express their opinions freely. They couldn't participate in determining Libya's course. Still, Libya had a more or less well-functioning government. It had executive, legislative, and judicial branches, albeit not of democratic standards. Libya was represented in international organizations such as the United Nations. It was also a candidate for becoming a member of the World Trade Organization. The international community judged Libya to be an autonomous state and treated it accordingly. Libya was an autonomous player on the international political stage.

At this point, again, the question arises: on what grounds do we count a state, even an undemocratic one, as autonomous? The answer is that a state counts as autonomous if it makes its own laws, determines its own institutional make up, independently selects its representatives, and participates as a unified entity in international relations. A state's autonomy is threatened from the outside when other states try to interfere with its internal affairs. It is threatened from within by the dissolution of internal order. Civil war, for example, can lead to a loss of autonomy because the state ceases to exist as a unified entity. This explanation makes no reference to the democratic constitution of a state. A dictatorship like Libya, for example, can make, without interference from the outside, its own laws, can determine its own institutions, and can participate as a unified entity in international relations.

The importance of this observation consists in the fact that the notion of self-directedness does not play the central role in understanding political autonomy. Let me expand on this somewhat more. To begin with, we need to keep in mind that the notion of self-directedness – when applied to persons – does not simply refer to the fact that a person acts intentionally. As I argued at some length in Chapter 5, the idea of self-directedness enters the picture precisely because we want to distinguish between mere intentional action and action that is backed up in some emphatic sense by what the agent really wants. Hence, when we try to find an analog to self-directedness in the political realm, we need to search for something like the state's self or its authentic standpoint. How is this supposed to work?

Let us assume for a moment that it makes sense to speak of a state's self or standpoint in the first place. My suggestion is that the most plausible account of political self-directedness is one that regards the democratic will of the people as the anchor for determining self-directedness. Based on this idea, a state is self-directed if and only if the democratic will of the people determines its institutions, actions, and so forth. If we conceive of a state's standpoint like this, we have a good analogy for an agent's standpoint because this allows us to distinguish between actions of the state which are backed up by its standpoint and actions that violate this standpoint. If we conceive of the democratic will of the people as determining the political self, we can count decisions, actions, or institutions that violate the democratic will as diminishing the state's self-directedness. In brief, then, a state is self-directed if and only if it is a democratic state whose actions have their source in the democratic will of its people.

Based on this assumption, we have to conclude that undemocratic states lack self-directedness. Proponents of self-directedness accounts of personal autonomy might try to avoid this result. Can we find another way to make sense of the idea that a state like Libya was self-directed? I think not. If we are at all inclined to use notions like self and non-self with respect to states, we are well advised to refer to the will of the citizens of the state. There is probably some room to maneuver here. But we would certainly twist the concept beyond usefulness if we were to allow that Libya's self is determined by Gaddafi's will. After all, Gaddafi's will was just one will among thousands of others. And there was nothing to it that made it apt to determine the true self of Libya. It just happened that he was in charge. But that by itself gave him no special authority to constitute Libya's self. Let us also remember that the self vs. non-self distinction gained traction in the debate about personal autonomy because people wanted to explain why some mental state that turns out to be stronger than all other desires and intentions still might not represent the agent's true self. The analogy to what in the personal realm is a mental state that usurps the agent is in the political realm a usurper at the head of the state. What this shows is that our assessment of a state's autonomy is not so much influenced by its self-directedness, but more by characteristics that are concerned with resolute agency.

This result comes hardly as a surprise when we remember the original usage of the notion of autonomy. There the idea of self-directedness does not play a role. Instead, the state's aptitude in dealing with external or internal opposition forms the

core of the concept. If we want to retain the analogy between the political and the personal realm, then we ought to acknowledge the essential role of resoluteness. In other words, against the background of the political understanding of autonomy, it is plausible to assume that autonomy encompasses more than just self-directedness.

The concept of personal autonomy that I am interested in retains the analogy to the political realm. That is, according to this concept, an agent's aptitude in dealing with external and internal opposition and conflict partly constitutes her autonomy. Mere self-directedness accounts of autonomy shed the analogy between political and personal autonomy. I think that it is rather counterintuitive because *prima facie* it is highly implausible to completely separate political and personal autonomy. Of course, we usually do not assume that these notions are identical. But it would be rather inelegant, given the history of both notions and their systematical treatment in philosophy, to treat them as if they were completely unrelated. The analogy to the political realm gives us a hint that a richer framework than that of pure self-directedness accounts of autonomy is required in order to explicate the concept of personal autonomy.

6.2 Achievement and Respect

In this section, I discuss a special value that is attached to autonomy as I understand it. I argue that this kind of value cannot be captured in terms of self-directedness. We need additional conceptual space to account for the value of autonomy. Unsurprisingly enough, my suggestion is, again, that the idea of resolute agency fits the bill. Here is the thought: autonomy is a value. In particular, it is a personal value. And it is a value that commands for respect. Does every person possess this value? No. As I argued at some length, natural autonomy comes in degrees and it is not an essential property of persons. Hence, a person might lack the value attached to it if she lacks natural autonomy. Why does it command respect? Autonomy commands respect because it is an achievement. The achievement involved in autonomous agency is that of doing something difficult, something that cannot be taken for granted. But self-directedness is not essentially tied to an achievement. Hence, autonomy goes beyond mere self-directedness. The additional dimension of autonomy is resoluteness, or so I will argue.

The achievement character stands out in all the examples that I have contemplated throughout the discussion. Think of Antigone, Socrates, Marie Curie, and Martin Luther, for example. For each and every one of them, sticking to their evaluational standpoint was a challenge. Marie Curie, for instance, managed to make an exceptional career for herself in science although she encountered a variety of severe obstacles. She moved to a different country in order to shape her life in accordance with her dream of becoming a scientist. And she endured the hardships women had to face in a situation dominated by men and their prejudice against the aptness of women trying to make a career. By standing her ground in these detrimental circumstances, she proved herself to be an exceptionally autonomous agent. Her case illustrates that leading one's life autonomously is a demanding business. It exemplifies the achievement character of autonomy. We see the value of this kind of autonomy, and we respect Marie Curie for her autonomous life.

When I say that autonomy has an achievement character, I want to highlight that being autonomous requires an effort from the agent. She needs to actively bring about or maintain her autonomy. In other words, leading one's life autonomously is something demanding. How demanding it is varies. For some people, it is more demanding than for others to act and live autonomously. Some people live in environments that present severe obstacles to an autonomous life. Some of us are luckier than others. But it is never the case that an agent acts autonomously as a matter of mere luck or happy coincidence. An autonomous action cannot just happen to the agent. An agent cannot be a passive bystander with respect to her own autonomous agency. In one reading, this is trivial because an action as such cannot be something that happens to the agent. Actions are of course activities. But the autonomous agent makes a special use of her agency. The agent determines whether she acts autonomously or not. And our attempts to act autonomously can always be unsuccessful. We might personally fail in our struggle for autonomy.

I said that autonomy is a value that commands our respect. The achievement character explains why autonomous agency commands this respect. We owe respect to people like Antigone, Socrates, or Marie Curie for standing their ground in the face of severe obstacles and opposition. Let me clarify the notion of respect at issue here. Sometimes we use the notion of respect slightly differently from how I use it in this context. Regarding the normative grounds of moral agency, we often speak of respect

for each other as a basic duty that does not have to be earned and that cannot be lost as long as one counts as a moral agent. This kind of respect is not contingent on some sort of effort or achievement on the side of the recipient. Quite the opposite: the notion is especially important if we deal with persons who lack certain skills. In this context, respect is importantly linked to protection.

In contrast to this, there exists a different sense of respect, which is pervasive in statements like, "Her way of dealing with the kids despite the fact that her husband just died deserves our respect," or, "We have to respect Tolstoy for writing such an epic novel." This kind of respect has something to do with how well people handle things. It is essentially connected to an achievement. Respect in this sense comes in degrees. Among other things, it correlates with the difficulties someone has to deal with in order to successfully achieve what she intends to achieve. When I say that autonomy commands our respect or that autonomous agency deserves our respect, I have this kind of achievement sense in mind.

This linkage to achievement distinguishes respect from some forms of admiration. We can admire someone for something that is beyond her control, like having an attractive face for example. But we cannot respect someone for having an attractive face because this is something that just happens to her. The notion of admiration might sometimes also be linked to achievement. When we admire someone for learning six languages, for example, the reason for our admiration is that learning six languages is quite an achievement. In a context like this, we can use the notions of respect and admiration interchangeably. However, to avoid misunderstandings, I will speak of respect when I refer to an attitude of esteeming someone for what she has achieved.

The self-directedness accounts cannot make sense of this value of autonomy because self-directedness is not necessarily an achievement. Consider the following example: a good angel has taken care of Lucky Linda. The work of the angel consists in making sure that Lucky Linda is self-directed. Without the silent work of the angel, many of Linda's decisions and actions would be determined by her non-self. What exactly does the angel's work consist in? Assuming for a moment that Frankfurt's account of self-directedness is true, we can imagine that the angel closely observes Linda's mental states and intervenes as soon as it detects that Linda is not wholehearted with regard to a certain situation. The intervention consists in modeling

Linda's mental states in a way that let her become wholehearted. For example, if Linda has conflicting higher-order volitions concerning her desire for going back to work because that would mean that she would spend much less time with her daughter, the angel would meddle with her higher-order volitions so that she either fully endorses or fully rejects her first-order desire. The angel also intervenes if it detects a conflict between competing first-order desires. In this case, it ensures that only those first-desires that Linda identifies with wholeheartedly become her will. Let us assume that Linda has a desire for spending time with her daughter and also a desire for going back to work. And let us further assume that she identifies wholeheartedly with the former and rejects the latter. In this case, the angel would secure that Linda acts on the desire for spending time with her daughter. We can stipulate that Linda would act on her desire for going back to work if the angel would not intervene. That is, without the intervention of the angel, a desire that Linda does not held wholeheartedly would become her will. We can imagine that Lucky Linda lacks the capacities to unify her will and that she sometimes is too weak to resist a desire that she rejects as really her own. The angel interferes so that she becomes wholehearted and acts accordingly. I assume that, as a consequence of the good angel's intervention, Lucky Linda is perfectly self-directed.

In this thought experiment, Lucky Linda passively receives her self-directedness like a gift. She fails in determining her own standpoint and seeing to it that she forms her intentions accordingly. For these reasons, her self-directedness is not an achievement of hers. It is a coincidence that she is self-directed. She exemplifies the right volitional structure without playing an active role in bringing it about. Hence, we have no grounds for respecting Lucky Linda for being self-directed. If she donates money, for example, because the angel made certain that the desire for donating money would outweigh all other desires, she played no active role in bringing about her will. Consequently, we don't owe her respect for behaving as she does, even though she is self-directed. This shows that mere self-directedness falls short of providing us with sufficient grounds for respect.

How can it be that we can count an agent as self-directed in Frankfurt's sense despite the fact that she is not responsible for bringing this about? The explanation for this is that Frankfurt does not investigate the causal relations between an agent's higher-order volitions and her will. Identification and endorsement do not imply any

causal relation. It is therefore possible that an agent's will does not have its causal roots in her higher-order volitions, but that it is brought about in some other way, by the intervention of the angel, for example. Still, the agent can endorse this will and thus count as self-directed within Frankfurt's framework. Now, if self-directedness were supposed to be sufficient for autonomy, we would get a different concept of autonomy because it would allow us to count an agent as autonomous who remains passive with respect to the formation of her will. A correction of the Frankfurtian picture that takes this problem into account consists in adding the requirement that the agent forms her will on the basis of her higher-order volitions. In other words, the agent must form this particular will because she identifies with it. Since this formation process can fail, we can rightly say that it is an achievement to have the volitional structure that realizes self-directedness.

Similar problems also arise in the Watsonian framework of autonomy as self-directedness. According to Watson, an agent is self-directed if and only if her intention is in accordance with her evaluational standpoint. He says that "when and only when agents' behaviour expresses their evaluations are they sources and 'authors' of (because they 'authorized') their behaviour."[288] Now it appears to be possible that an action expresses an agent's evaluational standpoint without having its causal roots in it. Let us imagine that Lucky Linda judges all things considered that she ought to donate money for children in poor countries. Let us imagine further that Lucky Linda fails to act on this judgment because she also has a stronger desire for spending her money on expensive paintings. If she buys herself a new painting, she is not self-directed, according to Watson, because this action is against her better judgment. Her action fails to express her evaluations. In a slightly different scenario, the good angel intervenes before Linda buys the painting. It models Linda's mental states so that she is most strongly motivated to donate the money. If she donates the money, she is self-directed because this expresses her all-things-considered judgment. However, as before, she is not actively bringing about the match between her judgment and her intention. Hence, it is not an achievement of hers that she acts in accordance with her-all-things considered judgment. And we do not owe her respect for this.

[288] Gary Watson (1987), 149.

Again, a solution to this problem consists in adding a causal requirement, namely, that an agent forms her intention because she judges that this is what she ought to do. With this requirement in place, only an agent who succeeds on the basis of her own effort to align her intentions and her evaluative standpoint would count as autonomous. In this case we would account for the achievement character of autonomous agency. In fact, I think that this additional requirement is amicable with Watson's idea that the agent is "authoring" her will. The thought experiment of Lucky Linda allows us to delineate different aspects of autonomous agency, namely, self-directedness and resoluteness. Watson's account does not give us a good explanation why Lucky Linda misses something as far as her autonomy is concerned. Her lack of self-control and her reliance on an external agent for modeling her psychological household do not count against her self-directedness, although they are problematic for her autonomy.

Let me finally address Michael Bratman's account of autonomy as self-directedness. In his notion of agential direction, Bratman explicitly acknowledges that self-directedness implies a causal relation between the mental states and the action. "As a first step we can say that for the agent to direct thinking and acting is for relevant attitudes that guide and control that thinking and action to have authority to speak for the agent – to have agential authority. […] When relevant attitudes with such agential control appropriately guide and control, the agent directs."[289] Guidance and control are causal relations. Building these requirements into one's account of self-directed agency is an improvement in comparison to Frankfurt's idea of identification and endorsement. But just as Watson, Bratman neglects to investigate resolute agency as a part of autonomy. If an agent like Lucky Linda comes to be self-directed by way of being modeled by a good angel, this does not speak against her autonomy. This overlooks the fact that autonomy has an achievement character that is violated in the case of Lucky Linda. In order to account for this achievement character, we need to introduce the notion of resolute agency.

To sum up: self-directedness accounts of autonomy cannot explain the special value of autonomy and why autonomy calls for our respect. The reason is that an account of self-directedness is not primarily concerned with the active role an agent has to play in securing her self-directedness in order to count as autonomous. Again,

[289] Michael R. Bratman (2007 a), 4.

autonomy has an achievement character. But such things as inner strength, persistence, and courage fall outside of the scope of an account of self-directedness. In order to bring these issues back into focus, we need to supplement an account of self-directedness with an account of resolute agency. The next section expands on this idea.

6.3 Courage and Self-Directedness

The considerations about achievement and respect that I have just presented support the claim that an understanding of self-directed agency falls short of accounting for those aspects of autonomy that have to do with inner strength, courage, self-control, persistence, and so forth. In this section, I present a further consideration that highlights the need for a notion of resolute agency in order to come to terms with autonomy as a whole. The basic idea is that there are differences in autonomy that are not due to self-directedness. These considerations also provide us with a clear description of what is missing, namely, an account of resolute agency.

Two agents who are equally self-directed can still differ in their degree of autonomy. Consider the following example: Weak Wendy chooses to study art and to live a life as a painter. Let us assume that Weak Wendy is self-directed in pursuing a career as an artist. Now imagine that Wendy lives in a social environment that is perfectly supportive. Her family, her friends, as well as everybody else support her in pursuing this career. Hence, Weak Wendy encounters no social pressure whatsoever. This is lucky for her because Wendy is indeed very weak: if she were confronted with any sort of social pressure she would give in to it. If her mother, for example, were to tell her that she should become a doctor instead of an artist, Wendy would pursue this different course of life. If the postman were to insist that she marry him and become a housewife, she would do as he wishes. In the actual situation, Weak Wendy is self-directed. However, she completely lacks the ability to resist social pressure. Lucky for her, she is not suffering from this lack because she lives a life without any social pressure. This example shows that it is conceptually possible that an agent who is *not* confronted with social pressure is self-directed although she lacks the ability to resist social pressure. That the agent lacks this ability is based on the truth of a counterfactual: if the agent were to encounter social pressure, she would comply with it. Let me call the ability to resist social pressure *courage*.

A self-directed agent who lacks courage is indeed possible. Assuming a Frankfurtian notion of self-directedness, Weak Wendy is, in the actual scenario, wholeheartedly pursuing a career as an artist. She desires to become an artist and she endorses this first-order desire, that is, she has a higher-order volition to act on this desire. Moreover, she is not troubled or internally divided about this in any way. This makes her wholehearted and hence self-directed, according to the hierarchical account. Her lack of courage does not show in the actual scenario since she does not encounter any situations that require courage. No one expects her to shape her life differently from what she wholeheartedly wants. But since the counterfactual is true of Wendy that she would give in to the demands of others if someone were to raise such demands, she is not courageous. There is nothing in the hierarchical account that excludes this possibility. According to Frankfurt, then, Weak Wendy is self-directed, even though she would act contrary to what she wholeheartedly wants if someone would demand this from her.

We can also describe an agent like Weak Wendy within Watson's framework. Let us assume that Weak Wendy acts, in the actual scenario, always in accordance with her evaluational standpoint. She judges, all things considered, that it would be best to pursue a career as an artist and thus forms the intention to do so. When she does all the things that are part of this project, she is self-directed. But again, there is nothing in Watson's evaluational account of self-directed agency that prevents us from stipulating that Weak Wendy is utterly uncourageous. She would violate her evaluational standpoint as soon as someone would demand her to shape her life differently. That is, she would still judge that she ought to pursue a career as an artist, but she would not find the courage to withstand the conflicting demands. She is too weak to stand her ground in cases of conflict.

Bratman's planning account of self-directed agency also leaves open the possibility of a self-directed agent who lacks courage. Weak Wendy could have a self-governing policy to treat her desire to pursue a career as an artist and all desires related to this as reason giving. In the actual scenario, her self-governing policies fully support the course she has chosen. But again, if she were to encounter any opposition, she would yield and do what is expected of her even if she stills judges, all things considered, that she should become an artist. This counterfactual makes it true that she lacks courage. We can even imagine that Wendy has a self-governing policy not to

treat her fear of opposing others as reason giving. That is, Wendy does not think that her uneasiness with social conflict gives her any reason to depart from her course. But when it actually comes to a situation of potential conflict, she avoids the conflict and violates her self-governing policies.

What are the consequences of these considerations for our understanding of autonomy? First, we should acknowledge that Weak Wendy is fully self-directed. No matter what account of self-directedness we might use, we can modify the example in such a way that Weak Wendy indeed is self-directed though lacking courage. Second, I think that it is intuitively plausible to say that Wendy is not fully autonomous. And, third, what accounts for her insufficient autonomy has to do with her lack of courage. Let me discuss this in some more detail.

The intuition that Wendy lacks in autonomy is backed up by the very basic idea that the autonomous agent has her own standpoint and is able to express it in her life. What Wendy lacks is the ability to express her own standpoint if she encounters conflicting expectations. This has a direct impact on her level of dispositional autonomy because dispositional autonomy consists in a set of dispositions and abilities to develop one's own standpoint and to stick to it when one encounters obstacles. We can illustrate this by comparing Weak Wendy to an agent who possesses a high degree of courage, like Martin Luther King Jr., for example. Martin Luther King Jr. is an exemplary case of a courageous agent: he stood up for what he thought to be right even though this caused trouble for him with the authorities and made him a target for hate, insults, and various attacks. He faced strong opposition; nonetheless, he expressed his opinions and engaged in a social movement he thought to be important. He was arrested and incarcerated, and still he was not intimidated into submission. Although he knew about the dire consequences of his actions, he stood up for his ideals. Let us assume that Martin Luther King Jr. is as self-directed as Weak Wendy. We still regard him as being more autonomous than Wendy. This shows that autonomy is not solely a matter of self-directedness. In addition to being self-directed, the autonomous agent possesses courage. Martin Luther King Jr. is exceptionally autonomous because he possesses this inner strength and courage. This corroborates the claim that the notion of self-directed agency is only a part of an adequate account of autonomy and that we need to acknowledge the importance of resolute agency for autonomy.

A possible objection to this line of reasoning consists in saying that Weak Wendy is not self-directed because self-directedness requires the agent to be minimally robust. If Weak Wendy is as easily discouraged from pursuing a career as an artist as the example suggests, then she lacks the minimal robustness that is necessary for true self-directedness. My reply to this is that this objection introduces a new notion of self-directedness that goes far beyond the notion of self-directedness as it has been developed by such authors as Frankfurt, Watson, and Bratman. I am quite sympathetic to the idea that something like robustness is an essential aspect of autonomy. In fact, this entire chapter is devoted to paving the way for a notion of resolute agency that figures in a comprehensive theory of personal autonomy. But I cannot see why we want to build this notion into the idea of self-directedness as it has been developed recently. And as I pointed out, according to the most important accounts of self-directedness, Weak Wendy can be understood as being self-directed.

Now one could grant that Weak Wendy is self-directed but insist that she completely lacks autonomy. If she is as incapable to resist others as I have suggested, then she lacks a necessary ability for counting as autonomous. Despite her self-directedness, she is not autonomous at all. But then it makes no sense to compare her to an autonomous agent like Martin Luther King Jr. This remark fails to see that I do not want to compare autonomous agents and look for a difference in self-directedness. My example illustrates just the reverse situation of two equally self-directed agents who differ in their autonomy. Having said that, let me point out that I do not argue that Weak Wendy really is wholly non-autonomous. Her lack of courage certainly infringes upon her autonomy. But I don't want to defend the stronger claim that she possesses no autonomy at all.

Is there really no way for Frankfurt, Watson, and Bratman to explain the difference in autonomy that we observe in Weak Wendy and Martin Luther King Jr.? Frankfurt, for example, could try to argue that Weak Wendy is not as wholehearted as Martin Luther King Jr. since she is so easily set on a new track. Somehow an agent like Martin Luther King Jr. seems to be psychologically more integrated. I think that such attempts are bound to fail. First, I cannot see why we should not count Wendy as being wholehearted in the actual scenario. I admit that she would not be wholehearted once she does what her mother or the postman wants her to do. But this counterfactual scenario does not change her mental structure in the actual scenario. Second, even if

Wendy were for whatever reason not wholehearted, there appears to be no conceptual inconsistency in imagining a wholehearted agent who lacks courage. And this is all I need for corroborating the claim that the notion of self-directedness lacks the resources to explain sufficiently all differences in levels of autonomy.

In a similar vein, we can reject attempts to account for the difference within the frameworks of Watson and Bratman. Without introducing the idea of resolute agency into these frameworks, we lack the resources to account for the differences in autonomy between two equally self-directed agents who differ in their courage. Of course, we will get different accounts of autonomy depending on what picture of self-directed agency we subscribe to. If we add resolute agency to the hierarchical account of Frankfurt, the result will differ from developing Bratman's approach in the direction of a more comprehensive theory of autonomy. But the question is not whether it is possible to accommodate these theories so that they include a systematical treatment of resolute agency. The question is whether they are able to account for issues concerning resolute agency as they stand. And the answer to this question is no.

I want to conclude this section with a general observation about why these self-directedness accounts of autonomy fail in this respect. What these theories have in common is that they focus on the actual situation. They primarily ask whether someone is locally autonomous. And they give the answer by identifying a kind of an actual psychological structure that supposedly grounds self-directedness. This focus leads to a negligence of the relevance of counterfactual scenarios for autonomy. And for this reason, the issues concerning dispositional autonomy are underemphasized. The notion of dispositional autonomy refers to the set of the agent's dispositions and abilities that allows her to prevail in conflicts and to overcome opposition. The dispositionally autonomous agent has what it takes to go her own way regardless of what others think about this. Systematically this can be explicated by taking into account two things: counterfactual scenarios and temporally extended parts of the agent's life. For the dispositionally autonomous agent, the counterfactual is true that she would prevail in particular situations marked by conflict and opposition. And the dispositionally autonomous agent proves herself to be autonomous not necessarily in every situation, but in the way she lives her life as a whole. The notion of resolute agency allows us to account for these aspects of dispositional autonomy. Resolute

agency describes the more or less stable properties of a person. And this provides us with the resources to ground counterfactuals and assessments of an agent's whole way of living.

6.4 Insufficient Self-Directedness and Autonomy

In this section, I discuss another example that highlights the necessity for supplementing self-directedness accounts with a notion of resolute agency. I want to emphasize at the outset that our paradigmatic examples of autonomous agents deal with self-directed agents. We think of people like Antigone, Socrates, and Marie Curie who are highly self-directed. But this does not mean that it only makes sense to think of autonomy with respect to highly self-directed agents. Indeed, if we focus on more mundane instances of autonomous agency we find cases in which agents who are inadequately self-directed still exhibit a high degree of autonomy. Their autonomy is due to their high level of resoluteness in what they are doing.

Let me start again by assuming a Frankfurtian picture of self-directedness. My aim is to describe an agent who is insufficiently wholehearted but resolute in standing her ground. The first part of this task consists in sketching an agent who has conflicting higher-order volitions pertaining to her will. Think, for example, of Troubled Tina who is deeply troubled about the question of whether she should pursue a career as a top manager. She desires to be responsible for big projects, thousands of people, and enormous amounts of money. She also desires to have an expensive lifestyle. These and other desires give her reasons to pursue a career as a top manager. However, Troubled Tina also desires to have a decent family life. She wants to see how her children grow up. And she loves to spend long hours talking to her husband. This is a rather typical dilemma for a lot of people: do I want to ascend the career ladder to ever-higher positions, or do I want to be a family person? Sometimes we are troubled by questions such as these and an answer does not always spring readily to mind. Troubled Tina is not wholeheartedly endorsing her desire to become a top manager and take on this role. Let us imagine that she has some higher-order volitions in support of this career and others that oppose it. This inner conflict does not render her incapable of moving forward. It is possible that her desire becomes her will, that is, that she acts on it without resolving the inner conflict. In this case, Troubled Tina becomes a highly ranked manager and starts to live up to the

responsibilities of this position while still being internally divided about whether she wants to act on this desire. Actually, I don't think that this example is far-fetched. Sometimes we need to make a decision and act on it even though we are not fully sure where we stand with respect to it. We might entertain strong doubts that this is the right path and continue to torment ourselves with the thought that we should do something else. This, at least, is the situation of Troubled Tina.

According to Frankfurt's account of self-directedness, Tina is not self-directed in fulfilling the demands of her new job because she is not wholehearted. This establishes the first part of the example. What we need in addition is that Tina acts resolutely. And indeed this is quite possible. Although her will does not reflect her undivided standpoint – simply because she lacks such an undivided standpoint – she still is able to express her will even when she encounters opposition. In other words, she is persistent and courageous in standing her ground even though she entertains doubts about her standpoint. For example, when she is convinced that a new product line will be a huge success, she fights for this new product line even if other powerful managers in her company disagree with her. If she doubts that a certain business strategy will succeed, she openly expresses her doubts to the executive board even if she knows that they love this strategy. And so forth. While she fights her colleagues, she acts on her desire to live up to the demands of her role as a responsible manager. But as we already know, she does not wholeheartedly endorse this desire. Hence, she is not self-directed, according to the hierarchical account. Still, it appears true that she has a certain degree of autonomy because she does not back off when her colleagues or her bosses oppose her.

In order to sharpen our understanding of this situation, let us compare Troubled Tina with Toni, who is also a top manager and is equally troubled about this role. Let us stipulate that Tina and Toni have identical first-order desires and higher-order volitions concerning the job as a manager. They are both lacking wholeheartedness about their career path. What distinguishes them is their level of persistence and courage. When Tina encounters opposition, she stands her ground. Toni, in contrast, feels intimidated by conflicting expectations and gives in to them easily. Hence, when Toni thinks that he has a good idea concerning the business strategy, he renounces it as soon as he believes that others might not like it. Moreover,

Toni often does what his colleagues tell him to do. This submissiveness makes Toni less autonomous than Tina.

This example raises an obvious question: is self-directedness even necessary for autonomy? If Tina lacks self-directedness but is nonetheless autonomous, then self-directedness seems not to be even a necessary condition for autonomy. But that directly contradicts my claim that self-directed agency is an essential aspect of autonomy. Hence, it would seem this example puts the whole argumentation in jeopardy. I will address this worry shortly. Let me first re-describe the example so that it fits the frameworks of Watson and Bratman.

According to Watson's evaluational account, an agent is self-directed if and only if she expresses her evaluational standpoint in her actions, whereby her evaluational standpoint is determined by her all-things-considered judgment about what she ought to do or what would be best to do. Given this approach towards self-directedness, we can imagine Troubled Tina as having trouble in making such a judgment. She has good reasons for both courses of action, that is, she sees the value in becoming a top manager as well as the value of spending more time with her family, therefore, she is unable to make a decision. In this case, fulfilling the role of a top manager is not backed up by her all-things-considered judgment. And this undermines her self-directedness, at least partly. As I said before, I think that it is a pervasive aspect of life that we sometimes follow a path while being uncertain that this is the right thing to do. If someone presses us, we have to admit that we clearly see the reasons for leaving this path. And we also have to admit that they might outweigh the reasons we have for staying on this path. But still we can walk this path. Again, this is the situation of Tina. Her lack of self-directedness is due to a lack of a definite standpoint regarding this issue. Still, she can react very differently to obstacles and opposition. Tina, by stipulation, is very determined not to let others interfere with her course. This distinguishes her from Toni who readily gives in to the demands of others. For this reason, Tina is more autonomous than Toni, although both lack self-directedness in equal degrees.

Someone might object that it is impossible to lack an evaluational standpoint. Sure, we might be unable to formulate our own standpoint in the heat of the moment. But this does not mean that we don't have one. After all, Watson points out that our evaluative standpoint is determined by those values and reasons that we accept "in a

cool and non-self-deceptive"[290] moment. And in such a moment of contemplation, we surely would come to a clear decision. In other words, our experience of being confronted with two or more options that cannot be ordered due to their value has to do with our epistemological limitations. But "in a cool and non-self-deceptive" moment, we would resolve the apparent dilemma easily. Hence, we always have a clearly defined evaluational standpoint.

I have two replies to this objection. First, I reject the underlying premises that there are no real dilemmas. Why shouldn't it be possible that two or more possibilities really are incommensurable and that an agent is not clearly on one side or the other?[291] It appears to me that this kind of situation is not only explainable by our limited epistemological capacities. But, second, even if it should turn out that we necessarily have a clearly defined evaluational standpoint, this would not speak, in principle, against the example. In this case, we would need to imagine that, unbeknownst to her, Tina's evaluational standpoint speaks in favor of spending more time with her family and quitting her job as a top manager. This would render her not self-directed and the example would still stand. I conclude that the evaluational account allows us to coherently conceive of an agent who lacks self-directedness, but who still shows autonomy in dealing with obstacles and opposition.

How can we accommodate the example so that it fits with Bratman's planning approach? According to Bratman, an agent is self-directed if and only if she is driven by a self-governing policy. A self-governing policy determines what desires to treat as reason giving. Now, it appears plausible that Troubled Tina views her desire to become a top manager as reason giving. Of course, we could deny this and describe her as viewing this desire as not reason giving. But, although possible, I don't want to go in this direction for two reasons. First, it is utterly implausible to model Tina like this. Her case would suddenly look like a paradigm example of weakness of the will. Second, this would make it much harder to understand her as being autonomous. We could still make a case for that, but I think that this would be too much of a detour for the point I am trying to make, namely, that there is more to autonomy than self-directedness.

[290] Gary Watson (1975), 215.
[291] For arguments concerning the existence of incommensurability compare: Joseph Raz (1986): *The Morality of Freedom* (Oxford: Oxford University Press).

I think that it is more fruitful to model Tina like we did in the Watson case. That is, Tina has conflicting self-governing policies and cannot find a clear position. This allows us to understand her as being insufficiently self-directed when she becomes a manager and fulfills this role. We might want to count her as being somewhat self-directed because, after all, she follows a particular self-governing policy. On the other hand, she surely is insufficiently self-directed because she lacks the inner cohesion that marks a unified standpoint. If we understand Tina in this way, we make use of the idea that self-directedness is a matter of degree. One can be more or less self-directed depending, metaphorically speaking, on how far away she is from her standpoint.

Given Bratman's framework, we understand Tina as insufficiently self-directed because she has conflicting self-governing policies and thus lacks a unified standpoint. At the same time, she possesses the inner strength and courage to prevail in cases of conflict. She does not shy away from conflicts with her colleagues or bosses. In other words, she shows autonomy in dealing with conflicting demands. This distinguishes her from, and makes her more autonomous than Toni, who bows down to the expectations of his bosses and colleagues, even if he disagrees with them.

At this point, I want to come back to the concern that we raised before, namely, that self-directedness appears not to be even a necessary condition for autonomy against the background of what I just discussed. There are two ways to describe agents like Troubled Tina. We could say that they completely lack self-directedness, or we allow for different levels of self-directedness and understand someone like Troubled Tina as being insufficiently self-directed. In other words, there are two basic options for understanding the self vs. non-self distinction. One could say that this is a binary distinction, according to which an agent is either fully self-directed or lacks self-directedness completely. Such an understanding appears to be natural when looking at Frankfurt's and Watson's accounts. In Frankfurt's hierarchical account, an agent either is wholehearted or not. In Watson's evaluational account, an agent either is acting in accordance with her all-things-considered judgment or not. In contrast to this binary reading, we could model the distinction along a continuum that has full self-directedness and total lack of self-directedness at its extreme points, but which also allows for intermediate levels of self-directedness. This latter possibility appears to be intuitively much more plausible. Agents can be

more or less in touch with what they really want. Troubled Tina surely is doing something that she values, at least to a certain extent. Even if it were to turn out that changing her course and becoming a family person would fit her desires and values even better than a life as a top manager, the latter also appeals to her and fits some of her basic values. She would miss her own standpoint much more if she were to quit her job, leave her family, and start a life as a professional gambler.

When I claimed that self-directed agency is an essential aspect of autonomy, I made the silent assumption that self-directedness is a matter of degree and not an all-or-nothing property. As I just explained this assumption fits our intuitions much better than the binary reading. Troubled Tina, for example, is not a marionette controlled by a nefarious neuroscientist with some fancy device. She is still determining her own actions. The fact that she is troubled and thus unsure of what course of action would be best does not render her completely irrational. And she does not only express the minimal rationality that also characterizes the akratic agent. In contrast to the akratic agent, Tina is not violating her own all-things-considered judgment. She is simply unable to formulate such an all-things-considered judgment. Of course, we can say that she lacks the emphatic self-directedness that marks the agent who has her own integrated standpoint and acts on it. But I think that we would lose descriptive power if we were to put her in the same box of non-self-directed agency in which compulsive, coerced, or manipulated agents belong.

I don't want to claim that the autonomous agent might lack self-directedness completely. My central claim is that autonomous agency requires a more or less unified agent who acts goal-directedly and for a reason. These requirements are not sufficient for self-directedness in the demanding sense. But they guarantee that the agent is at least minimally self-directed. Hence, when I claim that an agent like Troubled Tina is insufficiently self-directed, I mean that she fails to act on the basis of a clearly defined evaluational standpoint. At the same time, she does not completely violate her values, but acts according to what we might call a partial evaluational standpoint. But this observation alone does not give us an answer to the question how autonomous she is because autonomy is also constituted by an agent's resoluteness. Again, we see that self-directedness accounts of personal autonomy fall short.

One might object to my interpretation of the Troubled Tina example by pointing out that I give a wrong description of the action as she conceives of it. I base

my claim that she is insufficiently self-directed on the premise that her action is intentional under the description "this is what a top manager has to do." And my argument is that Tina is insufficiently self-directed with respect to the desire or judgment that grounds that action. The objector grants me for the sake of the argument that Tina is indeed insufficiently self-directed when it comes to fulfilling the role of a top manager. This, however, is not the action that we need to consider here. Let us focus on the situation in which Tina speaks up to her bosses. When she speaks up to her bosses, she is not intentionally performing the action "doing what a top manager has to do." For her, the action is intentional under the description "standing my ground" or "expressing my opinion" or something like that. And with respect to this action, she is fully self-directed, or so the objection goes. And the objector claims that we can re-describe all actions that I count as insufficiently self-directed but autonomous in this way, that is, all these actions are intentionally performed under a description that shows the agent to be fully self-directed.

Let me expand on this a little bit more. If the objection comes in a Frankfurtian guise, it runs as follows: although Troubled Tina is not wholeheartedly endorsing her desire to fulfill the demands of being a top manager, she wholeheartedly endorses her desire to stand her ground and express her opinion honestly. And this is the operative desire when she speaks up to her bosses. Hence, she is wholehearted. And there is no reason not to count her as being fully self-directed in the emphatic sense. Coming from Watson, we would need to understand Tina as judging all things considered that she ought to speak openly and honestly. This kind of action is backed up by her evaluational standpoint. Thus she is fully self-directed when she tells her bosses that their ideas are flawed. Finally, when we want to model the situation within Bratman's framework, we would understand Tina as having a self-governing policy to treat her desire to stand her ground and express her own opinion truthfully. Moreover, she is not in any way divided about being honest. Hence, when she goes into the meeting and makes herself and her position crystal clear, she is fully self-directed.

The aim of this objection is to undermine the claim that there is more to autonomous agency than self-directedness. In particular, we don't need to introduce the notion of resolute agency in order to fully account for autonomy. This is a serious challenge because it directly attacks the central claim of this study. However, I think that we have good grounds to reject it.

First, the objection exchanges the operative desire or motive. Even though this is possible, we need to see that this changes the example without good reason. Of course, in reality there might be all sorts of different reasons for which an agent can speak up to her bosses – or performs any other action for that matter. But there is nothing conceptually incoherent in understanding Troubled Tina exactly as I did. Without a further argument why we need to understand her differently, it simply begs the question to claim that she acts on a motive that makes her fully self-directed. Of course, we should notice that it is a difficult issue regarding how to individuate motives. And I didn't address the question how to identify the reasons for actions that are part of a larger plan. But these difficulties do not raise principle problems for my claim. So my first reply is that we can think of the desire to fulfill the role of a top manager as the operative one, and that this is sufficient to support my claim that resolute agency is not reducible to self-directedness.

A second and related reply acknowledges that we usually have problems to exactly identify for what reasons an agent acts because of our epistemological limitations. Usually there is a mixture of motives and considerations, and the agent might not even be aware of them. Hence, in reality, it is always possible that we falsely identify the agent's reasons for action. But with respect to the conceptual considerations that I have presented, these epistemological problems are irrelevant. In our thought experiment, we just need to stipulate the motivating desire and then describe the agent as being insufficiently self-directed. If we were to stipulate that this action is intentional for Tina under the description "telling the truth," then we need to describe her as not being fully behind this action. Maybe she read Nietzsche and came to believe that a moral attitude expresses some sort of inferiority. She decided that her desire to speak truthfully is not a desire she wants to act upon because she believes that this desire is a result and expression of her religious socialization, which she wants to leave behind. I admit that this is a little far-fetched, but again, it is not incoherent and it shows her as being insufficiently self-directed when she speaks up to her bosses.

We can also imagine that Troubled Tina encounters opposition when she considers taking the job as a top manager. As it were, she is troubled about this decision, and taking the job is not a fully self-directed action. But still, she could

decide to take the job despite the opposition that she faces. When she resists social pressure in order to become a top manager, she expresses her autonomy.

In order to sharpen the distinction between self-directed agency and resolute agency further, let us imagine an agent, say Karla, who just wants to live her quiet life without running into conflicts. We can imagine that Karla values blending in. Now, every once in a while, Karla cannot but raise her voice against other people's unjust behavior. When her boss mistreats one her colleagues, Karla thinks she should just be quiet because this would spare her a lot of trouble. In fact, standing silently on the side would be a fully self-directed action for her. But every so often, Karla finds herself taking sides with her colleague and opposing her boss. Even though she does not judge that this is what she should do, she expresses some autonomy by opposing her boss. Her autonomy is not grounded in self-directedness, but in resoluteness. She possesses the inner strength and courage to oppose other people and to prevail in social conflicts. These dispositions and abilities make her autonomous, even when she actualizes them in actions that are not self-directed in the emphatic sense.

Another objection that I need to deal with grants me that there is a sense in which Troubled Tina is autonomous, but that this is not the sense under which we value autonomy and aspire to be autonomous. Autonomy with insufficient self-directedness is not full-fledged autonomy. Someone who raises this objection could point to all the examples of autonomous agency that I mentioned in the last chapters. People like Antigone, Socrates, Martin Luther King Jr., and Marie Curie are paragons of autonomous agency. And something that we really respect about them is that they managed to be self-directed even though this was the hard way. We value self-directedness, and we value it even more if it has to be defended against people or circumstances that pose an obstacle. I concede that we usually have emphatically self-directed people in mind when we discuss the value of autonomy. But this does not amount to the claim that autonomy basically consists in being self-directed.

First, we respect someone like Troubled Tina for standing her ground even though she is insufficiently self-directed. And, as I argued in the last section, this respect is tied to the value of achieving something against opposition. In contrast to the submissive Toni, she certainly deserves respect for speaking her mind and holding her head high when she is faced with opposition. This in and of itself is a value. Indeed, because courage is a value, we try to raise our children in a way that they do

not feel intimidated by other people's expectations and can develop the ability to express their standpoint openly. Second, talk about full-fledged autonomy or some such thing forgets that autonomy is a matter of degree. People are more or less autonomous. And one reason for this is that they can be more or less self-directed.

Let me also mention that resolute agency and self-directed agency often come together. Hence it might appear to be the case that resolute agency has no intrinsic value. Resolute agents manage it more often than agents lacking resoluteness to be self-directed. For non-resolute people, self-directedness is only available as a matter of luck. Remember the case of Weak Wendy who is self-directed because her environment is perfectly supportive. Or think of Lucky Linda who relies on a good angel. In the real world, people usually face obstacles. Resoluteness helps them to deal with these obstacles. And for some people, this means that they live their lives more or less self-directedly. Please note that this explanation acknowledges the value of being self-directed. But, as I argued, the value of autonomy is not reducible to the value of self-directedness; it also entails the value of resoluteness.

Given the current state of the autonomy debate, it might seem rather surprising that I argue that even agents with minimal self-directedness can be autonomous. Let me give you a second example in support of it. Imagine two effectively brainwashed people, let's say Kim Kum-il and Ji Jun-il, who are two North Korean football players. They were subjected to severe manipulation during their upbringing in North Korea. They never had a chance to critically reflect on their values, to make their own decisions on important matters, or to get an idea of their range of alternative possibilities. Given that manipulation undermines self-directedness, both of them are insufficiently self-directed due the manipulation. However, it would be too hasty to deny them any autonomy. By comparing them with each other, we can identify some differences in their level of autonomy.

Imagine that both of them are playing football in the United States. Both of them are approached by journalists who want them to comment on the political situation in North Korea. As everybody else on the team, they have committed themselves to remaining silent about political matters. Due to an accumulation of unforeseen events, Kim and Jun find themselves stranded in the middle of New York. Soon they are beleaguered by journalists. And for the first time during their trip, they are without the help of an agent of the North Korean government, who usually blocks

all inappropriate questions. Kim resists any attempts made to squeeze some comments about politics out of him. Ji, however, is made of different stuff. He feels intimidated by the journalists and begins to tell them what they ask for. We can, of course, enrich this example in many ways so that it becomes more and more obvious that Kim does not bow down, whereas Ji readily breaks his commitments when he is pressed to do so. Kim certainly is the more autonomous agent. We might not value his political standpoint, but we surely owe him respect for acting in accordance with his beliefs. That he is insufficiently self-directed because of the lifelong manipulation by North Korean propaganda does not render him completely non-autonomous.

These considerations conclude my case for the claim that personal autonomy cannot be accounted for solely in terms of an agent's self-directedness. The considerations that I have discussed in this section point us in the direction of acknowledging resoluteness as an independent determinant of autonomy. On the one hand, this is somewhat surprising, given the fact that the contemporary debate was focused primarily on the idea of being self-directed and neglected a thorough investigation of resolute agency. On the other hand, for an unbiased look at the political roots of the notion as well as the paradigmatic examples of autonomous agents, it is hardly any surprise that resoluteness is an essential aspect of autonomy. It seems that we have lost track of the fact that autonomy is a conflict notion that is concerned with how an entity deals with real or possible conflicts. Once we acknowledge this again, it is quite obvious that we need to investigate in more detail what resolute agency consists in in order to develop an adequate account of autonomous agency. This is my major task for Chapter 7.

The considerations in this section give us good grounds on which to explicate autonomy as containing two dimensions, namely, self-directedness and resoluteness. Sometimes we focus more on one of these dimensions in our assessments of an agent's autonomy. But the autonomous agent is both self-directed as well as resolute. And neither of these two aspects can be completely missing without rendering the agent non-autonomous. The reason why I emphasize resolute agency in this chapter is that the autonomy debate has focused on self-directed agency. Let me conclude this section with some thoughts about the reason for this emphasis.

First, self-directedness is of major value for people. Quite often, a lack of self-directedness hinders the wellbeing of the person. Moreover, self-directedness and

authenticity often go hand in hand. And in the last century or so, we have developed a yearning for being authentic.[292] Hence, by and large, it is good to be self-directed. Second, self-directedness normally is a sign for other valuable qualities in a person. I have already pointed out that it appears to be a fact of life that you need to be autonomous in order to be self-directed, even though it is not a logical necessity. I suppose that this is the main reason why autonomy is often identified with self-directedness. There might also be a contingent connection between developing one's own autonomy and struggling for self-directedness. We encourage both at the same time. So we can partly explain the focus on self-directed agency by acknowledging, first, that it is valuable for us to be self-directed and that, second, self-directed agency and autonomous agency often go hand in hand. Because being self-directed and being autonomous are both valuable states that often occur together, they are easily confounded. Against this background, the inclination to account for autonomous agency in terms of self-directed agency is quite understandable.

As I pointed out in Chapter 5, the autonomy debate has focused on internal impediments to autonomy, i.e., compulsion, from the very beginning. This makes perfect sense if autonomy is primarily conceived of as an internal matter because compulsion is a problem that pertains to an agent's individual psychological make-up. And conceiving of autonomy in this way might appear to be legitimized, at least *prima facie*, if we interpret it along the lines of David Hume's notion of free will, as Frankfurt did. If we think that an agent is autonomous if she has the will she wants to have, then we naturally come to discuss what this psychological structure of wanting to want consists in. And in order to explicate this, we need examples of agency in which an agent does not want what she wants. Hence, compulsive agency comes into play. And since one natural way to understand compulsive agency is by distinguishing between what an agent does and what she really wants to do, the distinction between a self and non-self is only a step away.

What happened, once this distinction was in place, was that philosophers started to apply it to much more mundane and ordinary cases of agency. The distorted agency of an addict has been assimilated to ordinary cases of non-self-directed or inauthentic action, like doing an unsatisfying job, pursuing unfulfilling hobbies, or

[292] This "we" addresses people in Western democracies. In other cultures, authenticity might not be held in such a high esteem.

participating in social activities just because everybody does them. This assimilation gains traction as soon as notions like authenticity, identification, wholeheartedness, true will, real self and so on start to shape our understanding of autonomy. Now a lot of cases of non-compulsive agency count as non-autonomous since they fail to meet the requirements for self-directed agency. Suddenly, an ordinary person who is lacking in self-directedness is thrown into the same basket as the compulsive agent: the basket of non-autonomy.

6.5 Conclusion

This chapter provides us with a variety of considerations that corroborate a view of autonomy as an aptitude to prevail in different kinds of conflicts. It points to the direction of an understanding of autonomy that is more concerned with effective and resolute agency than with authentic and self-directed agency. The roots for this dimension of autonomy reach back to the political origins of the notion of autonomy. It is a pervasive feature of our intuitive grip on autonomy, as can be demonstrated by pointing to examples of such paragons of autonomous agency as Martin Luther King Jr. The focus on the self vs. non-self distinction, which marks the contemporary debate, leaves out the antagonistic element of autonomy. The autonomous agent is not only characterized by shaping her life in accordance with her own desires, beliefs, and values. What is even more essential is that she is shaping her life in a world that constantly presents different sorts of obstacles and hindrances that the agent needs to deal with. The autonomous agent is a resolute agent. Resolute agency is the topic of the next chapter.

7. Resolute Agency

In Chapter 5, I discussed the most important type of contemporary approaches towards autonomy, namely, accounts of autonomy that focus on self-directed agency. I argued that self-directedness is an important dimension of autonomy. Some of the central intuitions about the concept of autonomy and autonomous agency under consideration can be captured quite well with an account of self-directedness. The autonomous agent has her own standpoint. And this standpoint determines her actions. Some accounts of personal autonomy focus only on systematically explicating what self-directedness is. But as I have pointed out throughout the discussion, especially in Chapter 6, the concept of autonomy I have in mind is broader. An account of autonomy needs to be built around an understanding of what I call resolute agency.

There are, as far as I see it, two main reasons why we need to acknowledge that autonomous agency is partly constituted by resolute agency. First, as I have already emphasized, autonomy is a dispositional property of persons. Self-directedness is an actual property of persons. In order to account for autonomy as a dispositional property, we need an account of resolute agency. Second, autonomy is an antagonistic notion. It is concerned with how well people can deal with situations of conflict, broadly conceived. The notion of self-directedness does not entail an antagonistic element. Whether or not an agent is self-directed is, in principle, irrespective of her abilities to prevail in conflicts. The notion of resolute agency, in contrast, is essentially concerned with the agent's dispositions and abilities to master conflicts. Hence, in contrast to the notion of self-directedness, it allows us to account for the antagonistic dimension of autonomy.

We could see both of these shortcomings of self-directedness accounts of autonomy especially in the last chapter, in which I presented different considerations that all cast doubts on the idea that autonomy can be analyzed solely in terms of self-directed agency. I discussed examples of agents whose level of autonomy cannot be sufficiently captured in terms of self-directedness. Such agents as Weak Wendy and Lucky Linda lack resoluteness. They are insufficiently equipped to deal with difficulties and to prevail in situations of conflict. It is for this reason that their autonomy is diminished, even though both of them are self-directed. The other examples pointed in a similar direction. Troubled Tina, for example, strikes us as autonomous because she is a highly resolute agent. We have to admit that she is

insufficiently self-directed because her goals are not backed up by her unified evaluational standpoint. Nonetheless, when she goes after her goals, Troubled Tina is highly resolute, thereby proving herself to be autonomous. The autonomous agent shapes her life *against opposition* in accordance with her own desires, beliefs, and values. The dispositionally autonomous agent is the agent who possesses the dispositions and abilities to do just that: shaping her life against opposition.

In this chapter, I investigate resolute agency and its importance for autonomy in more detail. I begin with contrasting self-directed agency and resolute agency (7.1). I continue with discussing the idea that autonomy is an antagonistic notion (7.2). Sections 7.3 and 7.4 explicate the two dimensions of resolute agency, namely, persistence and courage. I deepen the understanding of resolute agency by applying Richard Holton's thoughts on resolutions and willpower to it (7.5). Against the emerging picture of resolute agency, I finally examine the paradigmatic cases of non-autonomy again (7.6).

7.1 Self-Directed Agency, Resolute Agency, and Dispositional Autonomy

In this section, I address differences between self-directed agency and resolute agency and discuss their relation to dispositional autonomy. The self-directed agent is characterized by expressing her own authentic standpoint. The resolute agent is characterized by surpassing difficulties and overcoming opposition. The self-directed agent expresses herself. The resolute agent prevails in conflicts. Whereas self-directedness is concerned with the question of whether or not an agent stands fully behind what she does, resoluteness is concerned with how well an agent manages all kinds of conflicts. Of course, both things often are intertwined. Expressing oneself can be a troublesome and conflict-laden issue and therefore requires resoluteness. Indeed, it is rather typical that resoluteness is required in order to act self-directedly. But while self-directedness describes a local property that an agent has with respect to certain actions, resoluteness typically refers to dispositional properties of agents. To say that an agent is resolute means that she possesses the abilities and dispositions that let her prevail in all kinds of conflicts.

By using the notion of resoluteness in its typical, dispositional reading, we can tackle a task that I formulated in Chapter 2. An adequate explication of the concept of

personal autonomy under consideration needs to account not only for local, but also for dispositional autonomy. Doing something autonomously is local autonomy. Local autonomy is always concerned with particular instances of autonomous behavior. Whereas local autonomy is concerned with a person's autonomy in particular instances of performing an action, forming an intention, making a judgment, and so forth, dispositional autonomy refers to the characteristics of a person as a whole. The question here is whether the agent is, in general, autonomous, and whether she lives her life autonomously. The notion of resolute agency accounts for dispositional autonomy. It describes those dispositions and abilities that constitute autonomy as a dispositional property of persons.

Is it really necessary to introduce the notion of resolute agency in order to account for dispositional autonomy? Why can't we just account for dispositional autonomy in terms of self-directed agency? Dispositional autonomy cannot be explained with the notion of self-directedness because being self-directed is not a disposition. Self-directedness is an actual property and not a dispositional one. We could try to explain what it might mean to be self-directed as a matter of disposition. But the answer to this question shifts into an explication of resolute agency. After all, an agent who is disposed to act self-directedly is an agent who is, for example, not easily intimidated by the demands of others, does not give up if she encounters obstacles, and possesses the self-control to resist temptation. These are the characteristics of resolute agency. Hence we need to introduce the notion of resolute agency into our framework of autonomy because this allows us to account for dispositional autonomy. A sole focus on self-directedness would leave a gap in our account of autonomy.

7.2 Autonomy – An Antagonistic Notion

Acknowledging that autonomy needs to be analyzed in terms of resolute agency also gives credit to the fact that autonomy is, according to the concept I explore, an antagonistic notion. By this I mean that autonomy is essentially concerned with conflicts. From the perspective of self-directedness accounts of autonomy, which dominate the current debate, real or possible conflict is only a contingent aspect of autonomy. How an agent deals with difficulties and conflicts in order to secure her self-directedness is not in the focus of the debate. As we have seen in the last chapter,

these approaches are even compatible with the possibility that an agent becomes self-directed without the need to endure in any kind of conflict. Weak Wendy was an example for this: she lives in a perfectly supportive environment and thus need not fight for her standpoint. She is self-directed in the actual scenario but would fail to be self-directed in every counterfactual scenario in which she is confronted with conflicts. The contemporary autonomy debate neglects to address the topic of resoluteness. Either it has a different concept of autonomy in mind or it gives us an incomplete picture of autonomy. Situations of conflict are the source of the relevance of autonomy.

As have I pointed out, to understand autonomy as an antagonistic notion is in agreement with its political origins. The political roots of the notion of autonomy are real or possible conflicts. The strong *polis*, that is, the *polis* that could defend itself and that could participate in military conflicts was conceived of as autonomous. Military aptitude was a defining feature of the autonomy of a *polis*. Moreover, the context in which autonomy was attributed to a *polis* was marked by the threat of being taken over by a superior power. This understanding of autonomy also explains why we paradigmatically think of social conflicts when we describe an autonomous agent. Martin Luther King Jr. proved himself to be an exceptionally autonomous agent by not letting himself be intimidated by the strong social opposition that he faced.

Situations of conflict typically are the reason why we contemplate questions of autonomy. Autonomy becomes an issue because we might fail to shape our lives on the basis of what we value and take to be important due to opposition and conflict. The autonomous agent is the author of her life. But this requires her to overcome obstacles and to endure in conflicts. In other words, she has to be a resolute agent.

Social conflicts paradigmatically shape our intuitive understanding of autonomy. Someone forces her will upon someone else, thereby violating the other person's autonomy. Someone does what is most important to her although others belittle her for this. But there exist other kinds of conflict, broadly conceived, which the autonomous agent can handle. The notion of resolute agency encompasses all of them. Generally, we can say that the resolute agent is characterized by abilities and dispositions to overcome obstacles and resist pressure. Specifically I suggest to distinguish two aspects of resolute agency, namely, persistence and courage. These two dimensions are in one way or another concerned with how an agent handles

certain kinds of detrimental influences. Persistence is concerned with abilities and dispositions to mobilize the necessary effort for realizing a certain plan. I also subsume basic proficiencies as an agent under the label of persistence. Persistence also encompasses what we might call self-control, that is, the agent's abilities and dispositions to modulate her own mental states in a way that furthers her prospects of being successful in whatever she attempts to do. Courage refers to a broad range of dispositions and abilities that are concerned with how the agent reacts to social influences. Of special importance are expectations and demands of others. The courageous agent remains, as far as her deliberations and actions are concerned, unimpressed by other people's expectations. Let me discuss persistence and courage in turn.

7.3 Persistence

I characterized resolute agency as an aptitude in dealing with conflicts, broadly conceived. The resolute agent prevails in conflicts and executes her intentions, even when she is confronted with detrimental influences. Persistence is an important aspect of resolute agency. In a first approximation, we can say that the notion of persistence refers, on the most general level, to a continued striving for achieving some goal, even though there is some cost involved. This might be slightly too general, though. Thinking of persistent agency conjures up pictures of an agent who has to mobilize a decent amount of effort in order to go after her goals. In either case, the agent's abilities and dispositions to put an effort into achieving her goals is the first aspect of persistence. The more effort an agent is able to mobilize, the more persistent she is.[293] This ability is especially important if a particular goal is hard to achieve.

Persistence is a property that describes how an agent goes after her goals. Calling someone persistent means, by and large, that she does not give up easily. Hence the first aspect of persistence, that is, the disposition to mobilize a lot of effort, aptly characterizes the persistent agent. But persistence in going after one's goals does not only have to do with summoning up more energy. We also think of the persistent agent as the agent who tries out alternative ways to reach her goals when she suffers a

[293] The importance of an agent's effort is highlighted in Robert Kane's libertarian account of free will. Although I reject the libertarian gist of Kane's approach I share his emphasis on the importance of effort in our agency. Robert Kane (1998): *The Significance of Free Will* (Oxford: Oxford University Press), Chapter 9.

drawback. Think of a doctor, for example, who tries to cure a patient. She tries the first treatment, but without success. She tries a second treatment, but again fails to cure the patient. Persistence consists in coming up with new ideas on how to achieve one's goal (in this case, new ways to treat the patient), and trying again and again. The disposition not only to try harder but also to search for alternative ways and to try out different routes to one's goal is a second aspect of persistence.

The persistent agent is determined to reach her goal. A lack of persistence comes in either of two ways: either an agent retains a goal but refrains from mobilizing the needed effort, or the agent drops the goal altogether because the road appears too rocky. The effect is identical in both cases: the agent refrains from going after the goal. Persistent agents, in contrast, summon up a lot of energy and seek out alternative ways for reaching their goals if it becomes necessary. This latter aspect proves her to be a flexible agent. Our everyday understanding of persistence does not imply a huge amount of flexibility on the side of the persistent agent. Sisyphus was certainly persistent in trying to place the stone on the top of the mountain, but not really flexible. However, I want to use the notion of persistence in a somewhat more technical sense that encompasses a general aptitude for successfully going after one's goals. Hence, I also count among the characteristics of persistence an ability to easily adapt to unexpected situations.

Persistence, as I use the notion here, also comprises dispositions and abilities that we might refer to as self-control, that is, abilities and dispositions to modulate one's emotional and motivational states in a way that allows the agent to go after her goals without being distracted by conflicting motivations or without being impaired in one's effectiveness due to detrimental emotions. An agent who goes after a certain goal might be tempted, for example, to let this goal fall to the wayside and take an easier route instead or get an immediate reward. The persistent agent stays on course. She has her goal in mind and is not easily led astray – neither by external obstacles nor by tempting alternatives. If she is confronted with temptation, she is able to exert self-control in order to reach her goal.

Here are some examples of persistent agency. Imagine that I want to lose some weight and my first plan is to stop eating chocolate. After a while, I realize that this is an insufficient plan: my weight does not drop. If I am persistent in reaching my goal, I employ a different strategy. I could try out a much more elaborate diet, or I could start

exercising. Regardless of whether I do the former or the latter (or maybe even both), it will take much more effort than I initially hoped. My success will depend crucially on my persistence. If my persistence is low, I would just shrug my shoulders, start eating chocolate again, and bemoan the world for its hardships. If I am a bit more persistent, I might mobilize myself to work out three times a week. Here is another example: if a persistent agent intends to open her own restaurant, she continues looking for possible locations, financial support, and qualified personal even when her initial efforts turn out to be rather unsuccessful. The persistent agent is prepared to put an effort into finding the right means for reaching her ends.

Persistence is typically needed for the completion of long-term projects like finishing one's degree, learning to play the piano, or running a marathon, because long-term projects require that the agent motivate herself not just once, but time and again, and often over longer periods of time. Moreover, the more complex a project is, the more unforeseen obstacles might arise. Persistence is required to continuously move forward even though the agent needs to deal with difficulties and obstacles. Even though it is natural to think initially about the completion of long-term projects when contemplating the value of persistence, the need for persistence is by no means restricted to them. If I want to speak to my boss in the next hour, I might need to knock at her door every five minutes till she finally has a small opening in her schedule. If I had been taken aback by the fact that she was busy the first three times, I might have missed the opportunity that was open to me the fourth time I tried. In other words, if I had been less persistent, I would have failed to talk to her. Take as another example a jammed door that you need to open. If you find that the door does not open easily and just stop trying, you show a lack of persistence. But if you try a little harder, maybe ram your shoulder into it, start kicking it, or ask someone to help, you are more persistent. Sometimes persistence shows itself in such small things as trying again after the line was busy the first time.

Another way to describe what I said about persistence is to say that the persistent agent is characterized by a high commitment towards reaching her goals. Being highly committed disposes the agent to mobilize additional effort, to search for alternative ways, and to flexibly adjust her behavior in order to reach her goals. We discussed in Chapter 4 that commitment towards goals is the hallmark of intentions. The persistent agent is quite firm in her intentions and is determined to execute them.

If we want to understand better how persistence is realized, we need to look at intentions because intentions realize agential control and a commitment towards action. We encountered a special kind of intentions when we discussed Bratman, namely, self-governing policies. Persistence is partly realized by a self-governing policy to refrain from treating one's desire for avoiding difficult courses of action as a strong reason. An agent with such a self-governing policy might have a desire for dropping a certain project when she encounters difficulties and the subsequent need for mobilizing an additional effort. But she does not view her desire as reason giving, or only as giving a minor reason.

Bratman's account allows us to describe a specific way in which persistence can be realized. A more detailed account of intentions that proves to be insightful in this context is Peter Gollwitzer's. Gollwitzer's account of intentions and how intentions support goal achievement is of special interest for us because it highlights the dimension of commitment in the functioning of intentions. Let me present his approach in some more detail.

At base, Gollwitzer distinguishes between two kinds of intentions:

> "we will make a distinction between goal intentions ('I intend to achieve x!') and implementation intentions ('I intend to initiate the goal-directed behavior x when situation y is encountered!'). Implementation intentions are seen in the service of goal intentions. The former commit the individual to specific plans as to when, where and how the latter are to be achieved."[294]

Goal intentions specify goals of an agent, for example, to become a medical doctor in Africa, to eat a piece of chocolate cake, or to write a novel. An important part of these goal intentions – apart from their representational content – is that they incorporate a commitment to really go after them.

The implementation intention, in contrast, specifies how I want to reach this goal. If I want to become a medical doctor in Africa, I have many alternative ways that would bring me nearer to completing this goal. I could study at different

[294] Peter M. Gollwitzer (1993), 142.

universities, specialize in different sub-disciplines, select different places in Africa, and so on and so forth. Implementation intentions specify for each of these alternatives the one I choose. In effect, the implementation intentions represent specific plans for reaching the goal.[295]

Using the terminology of Gollwitzer, we can say that the goal intentions of the persistent agent incorporate a particularly strong commitment towards reaching her goals. Gollwitzer observes that some people increase their efforts when they encounter difficulties. These people are not easily taken aback. "Instead, people may be found to stay in the field or try to intensify their efforts. When we observe such phenomena, we readily attest to the high commitment of the individuals involved."[296] And Gollwitzer tries to capture this high commitment, which I view as constitutive of persistent agency, in his account of intentions. For Gollwitzer, the highly committed agent is characterized by a certain kind of mind-frame that comes with the formation of intentions. He develops this idea on the basis of a four-phase model of goal achievement in which the transformation of mere wishes or desires into intentions is a crucial aspect.

> "It is suggested that the course of goal pursuit encompasses *four different, consecutive action phases* […]:
>
> 1. In the first action phase, called *predecisional*, people deliberate wishes and desires in an attempt to set priorities. To achieve this, the criteria of desirability and feasibility are employed. Selected wishes are highly desirable, but still feasible.
>
> 2. The subsequent *postdecisional*, but still *preactional* phase is characterized by efforts to promote the initiation of relevant actions via effective planning. The objective here is to get started with relevant actions, so that the realization of the selected wishes and desires is not put off.
>
> 3. Once relevant actions are initiated, the *actional* phase begins and the individual focuses on effectively achieving the desired outcomes.

[295] Compare also Peter M. Gollwitzer/Bernd Schaal (1998): 'Metacognition in Action: The Importance of Implementation Intentions', in: *Personality and Social Psychology Review* (2), 124-136.
[296] Peter M. Gollwitzer (1993), 148.

> 4. When these outcomes are finally attained, the postactional *evaluative* phase (where the individual compares what has been achieved with what was desired) is entered and the individual tries to find out whether further attempts at realizing the respective wish are worthwhile or even necessary."[297]

According to this model, the formation of an intention terminates the first phase. In the first phase, the agent has not yet decided what to strive for. Desires and wishes present themselves, and the agent needs to decide which of them she really wants to go for. Selecting suitable goals is the task that needs to be accomplished. Once the agent has decided on committing herself to a specific goal, she transitions into the second phase. This second phase is characterized by a search for specific ways of how to realize one's goals and deciding which course to pursue. Gollwitzer describes the transition from phase one to phase two as follows:

> "As long as people are torn between their various wishes and desires or are contemplating whether or not to pursue a certain wish or desire, they remain unable to get started on making them come true. This situation can be changed to the positive, however, by forming intentions such as 'I intend to pursue x!'. Resolutions of this kind terminate further deliberation as they result in a *commitment* to realize a wish or desire. What was characterized by velleity has now been transformed into a binding goal. We therefore refer to this type of intentions as *goal intention* […]."[298]

Goal intentions are essential for every endeavor. And the persistent agent is disposed to feel strongly committed towards the goals that she specifies in her goal intentions. This commitment is strengthened by forming "implementation intentions." Gollwitzer has shown in different studies that agents who form implementation intentions are much more successful in realizing their goals compared to agents who only form goal intentions. Implementation intentions increase the chances of success because the

[297] Peter M. Gollwitzer (1993), 149.
[298] Peter M. Gollwitzer (1993), 150.

agent readily uses the opportunities that present itself. Gollwitzer describes implementation intentions as follows:

> "One can easily resolve this type of conflict, however, by committing oneself as to *when*, *where*, and *how* implementation is to be started, as well as *what course* the subsequent goal pursuit is to take. This may be done by forming intentions, such as, 'I intend to initiate behavior x whenever the situational conditions y are met!' We call this intention an *implementation intention*, because it connects a certain goal-directed behavior with an anticipated situational context. Whereas the goal-intention described above commits the person to achieving a certain end state, implementation intentions commit the person to executing an intended goal-directed behavior once the specified situational context is encountered."[299]

The persistent agent focuses on information that is helpful for achieving her goals. She is not easily distracted. Moreover, when she encounters a situation that is apt for taking the next step towards completing one of her projects, she is prepared to act. With this, her chances of getting what she wants and thereby shaping her life in accordance with her own vision is much more likely. An agent enhances her persistence by forming implementation intentions.

Here are some general remarks about persistence. Persistence is distinct from one's level of expertise in a certain domain. Imagine two athletes, say tennis players, who play against each other. One of them might be the better player in terms of technique and tactical understanding of the game. If this player lacks persistence, however, the less skillful player might beat him because he pushes himself to his limits. The better, but less persistent player might shy away from the effort it would take to beat his opponent. Whether it is in a single match or in one's whole career, if one fails to summon one's best energies to be successful, one might never gain a success that would have otherwise been possible. This is of course true not only for athletic success, but for everything we strive for in our lives. Some goals stay out of our reach if we fail to mobilize significant effort to achieve them, even if we are, in principle, gifted enough to reach them.

[299] Peter M. Gollwitzer (1993), 151.

Is persistence necessary for achieving one's goals? An answer to this question depends on how high we set the bar for counting some exercise as expressing persistence. In general, it appears plausible to assume that some goals can be realized without any persistence. You need only a fairly limited amount of persistence, if any, for example, in order to finish your bowl of ice cream. Some tasks are so enjoyable in and of themselves that they require virtually no effort at all. Of course, I still need to do something, which always requires that I mobilize at least some energy. But it does not feel like a real effort. My mastery of some tasks might be so complete that doing them is easy, like a Sunday morning. I do not need persistence because completing these tasks seems to happen effortlessly. Hence, persistence is not a conceptually necessary condition for every kind of success. As a matter of fact, however, most people strive for a lot of things that they cannot reach without effort. And the level of mastery that allows us success in a certain domain without much actual effort is usually the result of practice. And practicing itself requires persistence in order to be successful.

Is dropping a goal or withdrawing one's commitment always a sign for insufficient persistence? The answer is no. An agent does not necessarily lack in persistence if she drops a goal of hers. It might be the case, for example, that she realizes that she wanted to reach two mutually exclusive goals and needs to drop one of them. Another good reason can be that the agent comes to believe that the value of the goal is not as high as the cost of accomplishing it. She wrongly believed it to be easily reachable and, hence, gave it a try. But once her assessment of the situation has changed and she comes to believe that the price outweighs the benefit of success, she has good reason to let go of this goal. Now, this latter case is a tricky one because one might argue that it is possible to construe every situation as one in which someone encounters difficulties and, as a consequence, one refrains from putting any more effort into reaching one's goal in terms of a new cost-benefit analysis that favors refraining. The difference lies in the weight with which an agent factors effort into her deliberation. The persistent agent views the fact that a certain course of action requires effort not in principle as a decisive reason against this course of action. For an agent lacking in persistence, the prospective need for exertion is already enough to refrain from heading into this direction. In between these two extremes lies a continuum of different weights someone assigns to a similar amount of effort. The more persistent an agent is, the less she views prospective effort as a reason that

speaks against a particular course of action. Hence, it is not required that the persistent agent continues going after some goal, no matter what kind of obstacles arise. Mobilizing additional effort is a costly business. It needs resources that might be better used in another project. In this case, the agent has good reason to let go of her goal.

Let me finally recap why persistence is important for autonomous agency. We said that the autonomous agent shapes her life in accordance with her own desires, beliefs, and values – and she does this against opposition. This latter qualification is an important one as I have argued extensively. Autonomy is an antagonistic notion. It is concerned with how an agent deals with conflicts. Persistence is a set of dispositions and abilities that allow the agent to retain her goals and execute her intentions, even if this requires her to mobilize a lot of effort or to try different approaches to achieve her goals. The autonomous agent is successful in achieving her goals. Otherwise she would hardly count as the author of her own life. The autonomous agent shapes her life rather than letting it be shaped by her circumstances, tradition, the expectations of others, and so forth. She successfully commits herself to certain goals. Having the ability to strongly commit oneself is an aspect of autonomy because a strong commitment defines a strong standpoint and facilitates successful goal achievement. Being a persistent agent is part of autonomy because this is necessary for achieving one's goals in a world that requires us to put an effort into getting what we want.

An agent with low persistence is more likely to fail in getting what she aspires to and in living the life she wants to live. Her perception of her possibilities is severely limited up to a point where she opts for certain courses of actions not because she views them as fulfilling, desirable, or good, but just because she shies away from the more attractive alternatives that require some effort to be realized. Of course, all of us are constrained by the circumstances we find ourselves in. Marie Curie lived in a country that did not permit women to study. She could not change this. However, that our options depend in a variety of ways on the context in which we live does not diminish our autonomy. The problem of the agent who lacks persistence is that she constrains herself without necessity. If someone aspires to become a doctor, but shies away from working towards that goal because she feels reluctant to expend the energy

that is required for successfully finishing her degree, then she is restricted by her own lack of persistence and not by the circumstances.

Persistence also plays an important role in explicating the idea that the autonomous agent is partly characterized by a form of self-creation or self-constitution. As I pointed out above, self-creation consists partly in determining the kind of person one is. The whole idea of giving shape to one's life has to do with how an agent can determine what kind of person she is. The character of a person describes aspects of her that are more or less stable. Persistence provides an agent with the means she needs to create stability in how she shapes her life. The autonomous agent does not only successfully complete long-term projects, but also reaches her "identity goals,"[300] as Gollwitzer dubs them. That is, she is successful in becoming the person she wants to be. Again persistence is of the utmost importance for that. A person who aspires to be a medical doctor and to help those in need in Africa encounters all sorts of obstacles on her way towards achieving her goal: her parents might object to this life plan, the degree might be quite challenging, and easier lifestyles readily suggest themselves. The autonomous agent can overcome these counterinfluences and lead the kind of life she wants to live and become the kind of person she wants to be, such as a medical doctor in Africa, for example. Persistence is a crucial factor in this long-term success. And the central determinant of an agent's persistence is her commitment to a certain goal. Being committed to reaching a certain goal helps the agent to mobilize the necessary effort and to find a way of getting there.

7.4 Courage

In Chapter 6, I introduced the notion of courage. Courage, in the technical sense that I employ in this discussion, refers to dispositions and abilities to prevail in social conflicts, broadly conceived. The courageous agent is not intimidated into submission by other people's demands. She is able to resist social pressure and to overcome social opposition if it is necessary to achieve her goals. I presented Weak Wendy as an example of an agent who completely lacks courage. She is disposed to follow other people's expectations if they differ from her own plans. Her inability to overcome social pressure is absolute. Martin Luther King Jr., in contrast, is an example of an

[300] Compare FN 71.

extremely courageous agent. He continued to fight for what he took to be important even though he had to face very strong social opposition.

This kind of courage is, as I see it, the most salient feature of autonomous agency. Autonomous agents stick to their path even when they are confronted with opposing demands and expectations. They do not bow down to the will of others and are not intimidated into submission when they encounter social opposition. It is therefore not a coincidence that paradigmatic examples for autonomous agency refer to people who possess a lot of courage, like Socrates, Mahatma Gandhi, or Martin Luther King Jr., or to instances of highly courageous actions, like publicly expressing an opinion that is opposed by the authorities as Copernicus did when he defended the heliocentric model, or Martin Luther when he pinned his theses to the door of a church in Wittenberg. These people are paragons of autonomy because of their exceptional courage in opposing other people's expectations. They were confronted with an extremely high amount of social opposition and still managed to cleave to their goals. For this reason, we naturally think of people like them when we are searching for examples of autonomous agency.

In the last section, I discussed persistent agency and argued that it is characterized by a set of dispositions and abilities to try hard, to try harder, and to try differently, if necessary. The basic idea is that the persistent agent continues to strive for her goals, even if this requires her to overcome obstacles and to circumvent difficulties. Against this background, we could view courage as an aspect of persistence. After all, it is natural to describe someone who continues to fight for her ideas against social opposition as an agent who is persistent in going after her goals. The reason why I still want to distinguish between persistence and courage is mainly a pragmatic one. As I just said, with respect to questions of autonomy, social conflicts are central. In order to do justice to this, I regard it as useful to introduce a notion that picks out those abilities and dispositions that are concerned with social conflicts.

Let me clarify again that I use the notion of courage in a technical way that slightly differs from our ordinary linguistic concept of courage. Usually we think of someone as courageous when she knows the dangers of a certain course of action and still chooses to follow it. And since this is also true for the paradigmatic cases of resisting social pressure, which are indeed marked by courage as we usually understand it, I use this notion here. The difference of my technical usage to the

everyday concepts lies, first, in an additional constraint: I restrict courage to contexts of social conflict. Usually courage applies to every situation of danger. Second, I broaden the everyday concept somewhat so that it also pertains to certain cases in which someone fearlessly accepts that others expect her to act differently without feeling any fear at all.[301]

The courageous agent is disposed to follow her own judgment instead of doing as other people expect her to do. Once she has formed an intention, she is disposed to retain and execute it, even if this brings her into conflict with other people. As before, we can understand the mental structure that realizes courage in terms of self-governing policies. The courageous agent has a self-governing policy to refrain from treating social pressure as reason giving. When she feels fear and a desire to give in to the social pressure, the self-governing policy instructs her not to base her deliberation about what she ought to do on these fears and desires. This strengthens her commitment towards her goals and makes it much more likely that she will successfully achieve them. And this is an essential aspect of the ability to shape one's life in accordance with one's own desires, beliefs, and values – against opposition.

What do I have in mind when I talk about social pressure? First and foremost, social pressure is concerned with what others expect me to do. Expectations are a necessary condition of social pressure. Some expectations are purely descriptive, like when you expect that it will rain today or that the billiard ball will go to the right when another ball hits it on the left. We also have these kinds of descriptive expectations with regard to other people. For example, we expect the weatherman to tell us about the weather, or we expect Tiger Woods to putt. These expectations are beliefs about what is likely to happen. In contrast to this, we also have normative expectations, as I want to call them. When I normatively expect something to happen, I think it should happen. When I normatively expect that you treat me with respect, I believe that you should treat me with respect. Social pressure has its source in normative expectations. I use the notion of an expectation as an umbrella term that refers, depending on the context, to such things as demands, orders, requests, or advice. The important feature is that someone thinks that I should behave in a certain way. From now on, when I speak about expectations, I mean normative expectations.

[301] One could argue that only the person who feels fear can be courageous. I don't want to decide this question here. As will become clearer, I use the notion of courage in a sense that views fearlessness as a result of courage.

Normative expectations are a necessary condition for social pressure, but they are not sufficient. Another necessary condition for social pressure is a prospective sanction. A normative expectation has to be paired with a prospective sanction for non-compliance in order to constitute social pressure. The sanction can be something as trivial as a look of disapproval or as harsh as being put into jail. It suffices that it is some sort of negative reaction. For our purposes, it is useful to distinguish furthermore between sanctions that are social in their nature from other kinds of sanctions. Social sanctions are reactions like disapproval, resentment, or ostracism. What they have in common is that the negative consequence directly regards the social status of a person and how she is respected by others. Other kinds of consequences, like being physically harmed or being fined, are not inherently concerned with social status. The border between social sanctions and other sanctions is not always clear-cut. Physical punishment is usually accompanied by disdain. And putting someone into jail typically has an effect on her social status. However, the immediate consequence does not consist in someone else's negative attitude towards me. A social sanction has, at its core, other people's negative attitudes towards the target of the sanction. Negative attitudes can range from disappointment to hate. Of course, how the negative attitude is expressed also determines the severity of the sanction. The differentiation between social sanctions and other kinds of sanctions allows us to define a narrow and a broad concept of social pressure. Social pressure, narrowly understood, is based only on social sanctions. The broader notion of social pressure also includes other kinds of sanctions. Accordingly, the notion of courage gets a narrower and a broader reading. Courage can mean that someone is not thrown off her course by prospective social sanctions. It can also mean that someone does not falter if she is threatened with such things as physical pain, imprisonment, or economic disadvantages.

So far I have said that social pressure consists in expectations that are accompanied by a prospective sanction. In order to see how well an agent is able to deal with social pressure, it is of course important to further assume that the agent is aware of these things. When a person lacks an understanding of normative expectations in the first place, let us say someone with a severe form of autism, then it is not a sign of courage if this person is not intimidated by others. Courage requires that the agent have at least a basic understanding of the sort of predicament she might find herself in if she does not comply with the expectations of others. This awareness

is not an additional necessary condition for social pressure, but it is a necessary condition for courageous agency. The complete ignorant is not courageous.

Let me say something about the phenomenology of social pressure. If someone perceives social pressure, she often experiences an inner conflict. The agent is aware of the fact that she prefers to not comply with the expectations of the other people. When she gives in, she has negative feelings like anger, fear, self-loathing, disappointment or some such things. In other words, she gives in reluctantly. However, there is also a sort of conformist behavior that creates no tension or inner struggle on the side of the agent. Some people are what I want to call happy conformists. The happy conformist is characterized by desiring what others desire and valuing what others value. She is not conformist out of fear, but because her desires and motives are mimicking what others regard as good or right. If a happy conformist is in the company of vegetarians, for example, she also becomes a vegetarian, and if she is in the company of hunters, she engages quite happily in hunting. This distinguishes her from the reluctant conformist who does what others expect of her because she wants to avoid the painful consequences. The reluctant conformist who despises hunting, for example, might still engage in hunting, if this is what others expect her to do, but she is not motivated by a desire to hunt, but by a desire to avoid being disesteemed by the others.

There is a second case of giving in to social pressure without reluctance or inner conflict. It is based on the possibility that one can have a motivation for compliance as such. Think of the soldier who does what his superior commands him to do – cleaning the latrines, for example. He can obey without reluctance and without developing a desire to clean the latrines. His obedience is grounded in a motivation to comply as such. It appears to be unnecessary to assume that he is primarily moved by a fear for painful consequences. Maybe it is something like second nature for him to promptly do what his superior tells him to do.

One might question that this case of blind compliance, although certainly conceptually possible, bears any empirical relevance. Isn't it the case that non-compliance with expectations is typically accompanied by being disesteemed? And as Adam Smith pointed out, "Nature, when she formed man for society, endowed him with an original desire to please, and an original aversion to offend his brethren. She taught him to feel pleasure in their favourable, and pain in their unfavourable regard.

She rendered their approbation most flattering and most agreeable to him for its own sake; and their disapprobation most mortifying and most offensive."[302] Hence, one might argue that we have a dominant motive to avoid disesteem, which always is part of the motivation to comply. I leave it open whether compliance without an avoidance motive is a realistic case or not. What I say about resilience and autonomy can account for this possibility. That is all I need.

Finally, there is a case that lies between complying reluctantly and complying without inner conflict. Social pressure can lead to what Richard Holton calls a "judgment shift."[303] As I will discuss in the next section, Holton refers to situations in which an agent's judgment is "corrupted" by temptation. This happens when the agent is confronted with temptation, reconsiders her intentions because of this temptation, and judges differently because of the temptation's influence. This analysis can be transferred to situations of social pressure. A judgment shift can also be triggered by social pressure. I might deliberate about the pros and cons of vegetarianism and conclude that, all things considered, I should continue eating meat. However, in a situation in which I am confronted with the disesteem by some colleagues of mine, I am so intimidated that I start to reconsider my intention to continue with my carnivorous lifestyle. As it happens, the social pressure leads me to view my reasons in a different light and to finally adopting the position of my colleagues, that is, I become a vegetarian myself. Sometimes, of course, it is very rational to reconsider one's intentions, and other people might help me to get a clearer and more justified judgment. However, when social pressure issues in a judgment shift, I first feel inclined to give in to the pressure and then search for a justification for doing so. Anticipating that I probably will give in to the social pressure, I come up with a justification for doing so. This helps me to avoid cognitive dissonance. Of course, this all can proceed more or less implicitly.

The courageous agent is not prone to judgment shifts when she finds herself in situations of social conflict. The cowardly agent might reason that it is the best thing to comply with social pressure. She initially feels a tension between her own intentions and plans on the one hand and the opposing demands on the other. But as

[302] Adam Smith (1982): *The Theory of Moral Sentiments*, (D.D. Raphael/ A.L. Macfie (eds.): Vol. I of the Glasgow Edition of the Works and Correspondence of Adam Smith; Indianapolis: Liberty Fund), 116.
[303] Richard Holton (2009): *Willing, Wanting, Waiting* (Oxford: Oxford University Press), 97.

soon as she feels that she will probably give in to the pressure, she starts to reconsider her intentions and finally forms a new judgment, according to which she ought to do what is expected of her. Holton describes this kind of judgment shift with respect to someone who gives in to temptation. But the same structure can be observed in situations of social conflict. Agents often justify their conformist behavior retrospectively.

Autonomy is diminished in all these cases of complying with social pressure because the agent is insufficiently courageous. Being deeply conformist is a paradigm example of a loss of autonomy.[304] The same holds for blind compliance. What is needed is courage. The courageous agent either lacks a desire for compliance or, if she has it, is able to control it. Moreover, she exemplifies stability in her commitment to a certain course of action.

Above I said that a person who is completely ignorant of the expectations and prospective sanctions that are directed at her does not express courage when she does not comply with social pressure. If the ignorant does not count as courageous because she simply is not aware of expectations and sanctions, what about the person who is so independent minded that she does not feel any social pressure? Let us imagine an agent who is so utterly self-confident and calm that she does not feel disapproval as a negative consequence. She knows that others expect her to behave in certain ways. And she also knows that these people will be displeased when she acts differently. But she does not feel the slightest concern about this. She stoically accepts that conflict is a part of life and has no inclination to make any rotten compromises. My intuition is that this person is extremely autonomous. In fact, her autonomy is so encompassing that she does not even feel troubled by prospective social sanctions. However, can we sensibly say that this person is courageous? After all, she does not need to overcome some fear of being punished simply because she does not fear it.

My take on these questions is that the highly independent agent who does not even feel social pressure has an exceptionally high degree of courage, and that this explains that she is not impressed in the least by other people's expectations. Of course, this explanation is based on the assumption that she is not in principle unable

[304] There are philosophers who think that autonomy is not necessarily undermined by conformism (Gerald Dworkin and John Christman, for example). However, as I pointed out in Chapter 2 this would be a different concept of autonomy from the one under investigation.

of being touched by expectations. If we imagine someone who is unable to feel and who, for this reason, does not feel social pressure, she is not courageous. But if someone who knows how disapproval feels is not impressed by social pressure, I regard her as courageous.

Courage allows an agent to withstand other people's expectations. Does this mean that the autonomous agent completely ignores other people's interests? Does someone forfeit her autonomy if she helps someone in need? And what about advice: is the autonomous agent deaf to other people's advice? Finally, is friendship and love out of reach for the autonomous agent because these kinds of relationships are partly constituted by certain norms and expectations? Autonomy would certainly lose much of its attractiveness and value if it were detrimental to all these things. Fortunately, though, autonomy is compatible with such things as helping other people, accepting advice, and caring for others. Courage enables you not to be intimidated by other people's expectations and prospective sanctions. It does not blind you to other people's needs nor does it deafen you to the sensible things they have to say. Autonomous agency is undermined if one gives in to social pressure. That is, if my reason for action is that other people expect me to act in a certain way and I fear the consequences of non-compliance, I am giving in to the social pressure. But if I think that I have a good reason for this action apart from the other people's demands, I am not giving in to social pressure, even though my action is in accordance with the direction the social pressure tries to enforce on me.

Here is an example: if my reason for helping a child that had an accident is that other people watch and I fear that they will chide me if I refrain from helping, I am not acting autonomously. However, if my reason for helping in these circumstances is that the child needs help, my autonomy is not impaired. Similarly, if my reason for following someone else's advice is that I come to judge that I have indeed good reason to do as advised, I am not necessarily impaired in my autonomy. However, if I do as advised because I fear disappointing the other person if I do something else, I am giving in to social pressure and hence diminish my autonomy. In other words, that someone acts in accordance with other people's expectations does not necessarily undermine her autonomy. The important question is why she does it. If someone's decisive reason for action is that she stands under great social pressure, she is acting non-autonomously. If she has other reasons, she still can be autonomous. It is

not a necessary aspect of autonomous agency to oppose other people's expectations no matter what.

Let me mention finally that we can make sense of the idea that autonomy has a special value that calls for our respect when we acknowledge courage as a constituent of autonomy. Courageous agency is marked by an exceptional effort and is a prime example for an achievement if the agent is successful. Resisting other people's demands and following through with one's own ideas typically is an effortful business. If an agent lacks courage, the fact that she is self-directed does not command our respect. The special kind of value which I described in Chapter 6 requires resoluteness and, in particular, courage.

7.5 Richard Holton on Resolute Agency

To deepen our understanding of resolute agency further, I want to discuss Richard Holton's approach towards strong willed action. In the center of this approach stands the notion of a special kind of intentions, which Holton calls resolutions. I will use Holton's concept of resolutions and Holton's willpower account of strength of will to explicate what resolute agency consists in in more detail. Resolute agency is marked, first, by a particular strong commitment towards certain kinds of actions or certain values in deciding what to do. In addition, the resolute agent is able to strengthen her will so that she does not falter in her commitment. We find both of these aspects of resolute agency explicated in Holton's account.

All intentions are characterized by a commitment towards action. We discussed this feature of intentions in Chapter 4. This general commitment towards action consists in being settled upon a certain course of action. That is, the agent is not ambivalent with respect to whether or not she aims at performing a particular kind of action. She is decided. Holton points out that some intentions embed a special commitment that goes beyond this general commitment. Some intentions possess the special commitment to retain the intention even if the agent is confronted with certain inclinations to drop it. Holton refers to this class of intentions as "resolutions." The special thing about resolutions is that they are "*contrary inclination defeating*."[305] For example, my intention to get up tomorrow morning at six and go out for a run would

[305] Richard Holton (2009), 77.

be a resolution if I would add that I don't want to drop this intention just because I feel tired or have a desire for reading the newspaper or dislike the weather. As a resolution, my intention shields itself against contrary influences that I anticipate in the future. In contrast to simple intentions, resolutions are formed with an eye on desires and inclinations that the agent anticipates as possible defeaters of this very intention. The resolution entails a commitment to uphold the intention even if those desires and inclinations start to pull the agent in a different direction. "Resolutions serve to overcome the desires or beliefs that the agent fears they will form by the time they come to act, desires or beliefs that will inhibit them from acting as they now plan."[306]

Resolutions facilitate resolute agency because their function is to let the agent retain her intentions and execute them in adversary conditions. Hence, Holton's notion of a resolution allows us to explicate part of the concept of resolute agency in more detail. The resolute agent is an agent who possesses the ability to form resolutions and is disposed to form them whenever she intends to do something of importance. By forming resolutions, an agent increases her chances of successfully reaching her goals because she shields herself against some kind of contrary influences. An agent's persistence as well as courage can both be realized by resolutions.

Resolutions, as Holton conceives of them, are instruments of hierarchical control because they are partly constituted by higher order intentions:

> "At the most intellectual level, resolutions can be seen as involving both an intention to engage in a certain action, and a further intention not to let that intention be deflected. Understood in this way they involve a conjunction of two simpler intentions, one first-order and one second-order (i.e., an intention about an intention). So, when I resolve to give up smoking, I form an intention to give up, and along with it I form a second-order intention not to let that intention be deflected."[307]

[306] Richard Holton (2009), 77.
[307] Richard Holton (2009), 11.

It is clear that an agent who resolves to do something is better prepared to actually do it because she both reflects beforehand on possible obstacles that might lead her astray and commits herself not to count these obstacles as giving reasons against her action. She does not want to be moved by them. By preparing herself like this, she enhances her chances of successfully executing her intention and shaping her life according to her own desires, beliefs, and values. An agent who is well equipped to reach her goals against opposition is dispositionally autonomous.

Holton sharpens and extends this account of resolutions by using it as a key notion in an approach of weakness of will and strength of will. I want to introduce this account in some detail because it allows us to explicate more aspects of resolute agency.

Let us first focus on weakness of will. Traditionally weakness of will is understood in terms of akrasia, that is, as "action voluntary undertaken against one's best judgment."[308] At the center of this understanding lies the idea of the rational agent who is able to contemplate reasons and to make a judgment about what she ought to do, all things considered. In akratic action, the agent forms such a judgment, but fails to act accordingly. Instead, she performs an action that is against her better judgment. Holton rejects the claim that weakness of will just is akrasia. "I shall develop the idea that the central cases of weakness of will are best characterized not as cases in which people act against their better judgment, but as cases in which they fail to act on their intentions."[309] Holton argues that the weak-willed agent is characterized by an over-ready intention revision. "Weakness of will arises […] when agents are too ready to reconsider their intentions."[310] This over-readiness might or might not be accompanied by akrasia. Before we examine Holton's distinction between akrasia and weakness of the will, let me clarify the notion of weakness of will a little bit further.

As a first point of clarification, Holton points out that not every revision of an intention is a case of weakness of will. The central question is why an agent revised an intention. Weakness of will consists in revising an intention on the basis of a deliberation that the agent should have refrained from performing. In Holton's words,

[308] Richard Holton (2009), 72.
[309] Richard Holton (2009), 70.
[310] Richard Holton (2009), 71.

"actors show weakness of will when they revise an intention as a result of a reconsideration that they should not have performed; that is, when their reconsideration exhibits tendencies that it is not reasonable for the agent to have."[311] Holton does not pause to systematically investigate how to explicate what those tendencies are. But he hints at the direction by giving some examples of reasonable maxims of reconsideration such as:

> "- it is reasonable to have a tendency to reconsider intentions if one believes that circumstances have changed in such a way that they defeat the purpose of having the intention;
>
> - it is reasonable to have a tendency to reconsider intentions if one believes that they can no longer be carried out;
>
> - it is reasonable to have a tendency to reconsider intentions if one believes that they will lead one to great suffering when that suffering was not envisaged at the time of forming the intention […]"[312]

I follow Holton in assuming that we have a good intuitive grasp on what kind of tendencies of intention reconsideration are reasonable to have. In this context, Holton also speaks of "the norms of the skill of managing one's intentions,"[313] and highlights that the account of weakness of the will contains two parts, a descriptive and a normative one:

> "The account offered here employs both a descriptive and a normative element. To display weakness of will the agent must have formed a resolution that they then revise in response to the very inclinations that it was supposed to defeat; that is the descriptive element. And their revision must be something that, by the standards of a good intender, they should not have done; that is the normative element."[314]

[311] Richard Holton (2009), 75.
[312] Richard Holton (2009), 75.
[313] Richard Holton (2009), 73.
[314] Richard Holton (2009), 89.

Spelling out the normative element is a challenge, because there appears to be no complete consensus on what kind of tendencies for intention revision it is rational to have. But although we might lack a clear-cut systematical account of this normative element, we have some good intuitions about what might be reasonable and what would not. Holton's rules of thumb that I just cited seem to be uncontroversial.

Another clarification is meant to contrast weakness of will with caprice. The capricious agent often changes her intentions, jumping from one alternative to the next and back again. But although this also is a case of over-ready intention revision, it is somewhat implausible to subsume it under the heading of weak will agency. Thus Holton introduces the following distinction: "If someone over-readily revises a resolution, that is weakness of the will; if they over-readily revise a simple intention, that is caprice."[315] Summing these ideas up, Holton formulates the following definition of weakness of will: "weakness of will is unreasonable revision of a contrary inclination defeating intention (a resolution) in response to the pressure of those very inclinations."[316]

The weak-willed agent, then, is an agent who drops an intention because of contrary influences that she excluded as reasons for letting go of the intention. In other words, the weak-willed agent is defeated by the very obstacles that she was resolved to overcome. When I form the intention to have salad for lunch, and when I further form the intention to retain this intention even if I crave pizza once I order lunch, then I prove myself to be weak willed when I order pizza for lunch just because I feel a craving for pizza. My resolution was not to be led astray by a desire for pizza. When I give in to this desire nonetheless, I behave in a weak-willed fashion.

Holton argues that this understanding of weakness of will is much more in line with our common understanding. Everyday characterizations are things like "weak willed people are irresolute; they don't persist in their intentions; they are too easily deflected from the path they have chosen."[317] This understanding differs from the philosophical orthodoxy that equates weakness of will with akrasia. Let me just

[315] Richard Holton (2009), 77.
[316] Richard Holton (2009), 78.
[317] Richard Holton (2009), 70.

mention some of the considerations that speak in favor of Holton's account of weakness of will.

One important advantage is that it allows us to account for cases of weak-willed agency in which the agent is unable to judge which alternative course of action would be best. She sees several alternatives that all have something attractive, but she cannot decide which one is the best. Forming an intention nonetheless is an important ability in order to avoid paralysis.[318] In this case, the intention is not backed up by an all-things-considered judgment about what ought to be done. Hence, if the agent drops this intention and pursues one of the other attractive courses of action, she is not akratic. But given that we spell out the details right, she certainly is weak willed because she unreasonably revises her intention. Think about the young man Sartre asks us to envision.[319] He is drawn between joining the resistance against the Nazis or taking care of his ill mother. He cannot do both. But neither is he able to judge one of these things to be more important than the other. Let us imagine that he finally decides to join the resistance – not because this is what he ought to do more than caring for his mother. But he needs to do something. He forms the intention to join the resistance. Moreover, since he knows how hard he struggled to make a decision, he also intends not to be led astray by his desire to be a good son and stay with his mother. The next morning he packs his things and tells his mother goodbye. But seeing her suffering reminds him that, as a good son, he should stay with her. And so he unpacks again and tells her that he will not leave. It is not the case that he now judges that it would be best to stay with his mother, better than joining the resistance. Hence, it is neither akratic to leave nor to stay. But after forming the resolution to join the resistance, he proves himself to be weak willed when he stays with his mother. Sartre's example is a rather dramatic one. But the predicament is a common one: sometimes we cannot decide which course of action would be best. We still need to make a decision, however. And once we have resolved to pursue a certain course, backing off, even though we did not acquire new relevant information, becomes a sign of weakness of will.

Another consideration that speaks in favor of Holton's account is that it avoids the following problem: let us imagine an agent who decided to do something, let's say

[318] "Often we have to act in the absence of a judgment of what is best. That is way we need a capacity to make choices that is independent of our judgments." Richard Holton (2009), x.
[319] Jean Paul Sartre (1948): *Existentialism and Humanism* (London: Methuen).

to quit smoking. Let us imagine further that this agent thought the matter through and has seriously contemplated all the good reasons for quitting. Against this background, she finally judges that she ought to quit. She throws away her last cigarettes, makes herself a cup of coffee, and enjoys her new, healthy life with a smile. An hour later a friend drops by, they start chatting, and after a while the friends lights himself a cigarette and asks her whether she also wants one. She remembers her resolution to quit. But in this very situation she thinks that it would be really nice to have a cigarette with this good friend. After a moment, she judges that this is really what she ought to do: she ought to take the offered cigarette. And since she is not akratic, she acts accordingly – she starts to smoke again.

This is a blatant example for weakness of will. One might want to object that people can change their minds. And this is what happened here. I agree that we need to allow for changes of mind that are not weak willed. After all, people acquire new information, the situation evolves, and so forth. But it appears to let people too easily off the hook from weakness of will when they can avoid this charge just by forming the judgment that it would be best to do something that they a moment ago resolved to avoid. Holton's account allows us to systematically distinguish between those changes of mind that are weak willed and those that are not. A change of mind is weak willed if the agent undergoes it for the very reasons that she prepared against.

Weakness of will is a failure to uphold one's intention in the face of temptation. The weak-willed agent gives in to temptation, thereby thwarting her plans. How does temptation work? From the akratic view of temptation, the agent who gives in to temptation acts against her better judgment. In contrast to this, Holton argues that temptation typically leads us to accommodate our judgments, that is, when we give in to temptation, we align our judgments with our desires. "I argue that temptation frequently works not simply by *overcoming* one's better judgment, but by *corrupting* one's judgment."[320] Holton dubs this process "*judgment shift*."[321] A judgment shift occurs as a means to avoid inconsistency. Holton refers to cognitive dissonance theory in order to strengthen his claim that temptation typically works by eliciting a judgment shift. "In general we work very hard to ensure that the picture we have of ourselves is coherent: that it is not 'dissonant'."[322] Hence, when we are

[320] Richard Holton (2009), 97.
[321] Richard Holton (2009), 97.
[322] Richard Holton (2009), 100.

confronted with a tempting alternative to our chosen course of action, we might revise our judgment on what we ought to do so that the tempting alternative suddenly presents itself as the best alternative. This process lies at the heart of temptation: "in many standard cases of temptation we get judgment shift: where the options are judged as close, judgments are revised to bring them into accord with desires, rather than desires being revised to bring them into accord with judgments."[323]

Holton points out that temptation is often distinguished from compulsive behavior, for example, addiction. Whereas the tempted agent retains the ability to resist the temptation, the compulsive agent lacks the power to resist her compulsion. "In the standard cases one succumbs to a temptation whilst retaining the ability to resist. As a result one maintains one's agency. In contrast, in cases of addiction – and perhaps in other cases like obsessive-compulsive disorder and kleptomania – one's ability to resist, and hence one's agency, is lost."[324] Holton disagrees with this standard picture of addiction by claiming, "that addiction should not be *defined* as a state in which an agent cannot resist temptation."[325] Of course, addiction often has the consequence that the addicted agent gives in to temptation. But "[w]hat is distinctive about addiction is that it involves a specific form of decoupling between, on the one hand, one's judgments, and, on the other, one's desires."[326] This decoupling leads the agent to continuously desire something, let's say taking heroin, even though she judges that this will give her no pleasure. "Normally if one judges that one would get no pleasure from satisfying a desire, its force is automatically undermined; a similar effect arises when one judges that one ought not to satisfy it. Addiction undermines these links."[327]

Holton grants that this understanding of ordinary temptation and addiction still views them as distinctive phenomena. But he hastens to add that on his account they bear more similarities with each other than they do in the traditional understanding. "So whilst I argue that ordinary temptation does differ from addiction – the first involves judgment shift, the other involves a decoupling of judgment from desire – one consequence of my account is to bring them more close together than in the

[323] Richard Holton (2009), 110.
[324] Richard Holton (2009), 97.
[325] Richard Holton (2009), 98.
[326] Richard Holton (2009), 98.
[327] Richard Holton (2009), 98.

standard philosophical picture."[328] Moreover, they are not completely independent from each other. "In particular, if decoupling comes in degrees, which looks very likely, then it is plausible that very many cases of ordinary temptation will involve it to some extent."[329]

An example for a judgment shift is the person who decides to quit smoking – and changes her judgment at the very next opportunity to get a cigarette because she has this strong desire to smoke. Heavily addicted persons exhibit the decoupling of desire and judgment. They crave the drug even though they believe that it won't bring any pleasure. "They need not like the substances to which they are addicted: they need take no pleasure in getting them, nor in the prospect of getting them."[330] This latter observation leads Holton to distinguish between a "liking system" and a "wanting system."[331] Liking something has to do with perceiving it as pleasurable. "Wanting something, in contrast, can be identified by its impact on being motivated to get the thing."[332] The addicted patient wants something despite the fact that she doesn't like it.

Although addiction in its extreme form is comparatively rare, the decoupling of desire and judgment, and also the decoupling of the liking system and the wanting system, can be found to lesser degrees also in healthy persons.[333] Judgment shift appears to be a very widespread phenomenon. Holton even claims that it occurs in most cases of giving in to temptation. "When we succumb to temptation we tend to judge that that is the best thing to do."[334]

Weakness of will undermines autonomy because it thwarts the agent's attempts to express her evaluative standpoint. The weak-willed agent is impaired in her self-directedness because she acts contrary to her practical identity. Resolute agents possess a strong will. They pursue their projects even if they encounter obstacles and opposition. Think of Marie Curie, for example, who refused to let go of

[328] Richard Holton (2009), 98.
[329] Richard Holton (2009), 98.
[330] Richard Holton (2009), 104.
[331] Richard Holton (2009), 104.
[332] Richard Holton (2009), 104.
[333] A decoupling of desire and judgment is not identical to a tension between the liking system and the wanting system. Although it is plausible to identify desiring and liking something, the same is not true for judging and wanting. An agent who judges that she ought to x does not necessarily want to x or want it more than anything else.
[334] Richard Holton (2009), 100.

her goal to become a scientist even though the odds were against her. Or think of Martin Luther King Jr. who continued his participation in the civil rights movement even though people opposed him very strongly.

The opposite of weakness of will is strength of will. Holton posits that there exists a distinct faculty of willpower that underlies strength of will. "My claim is that willpower is substantial; it is at least a skill and perhaps a self-standing faculty, the exercise of which causally explains our ability to stick to a resolution."[335] This assumption provides the basis for what Holton dubs the "willpower account" of strength of will:

> "Action is not determined just by the agent's beliefs, desires, and intentions. In addition willpower plays an independent contributory role. Agents whose willpower is strong can stick by their resolutions in the face of strong contrary desires; agents whose willpower is weak readily abandon their resolutions even when the contrary desire is relatively weak."[336]

The willpower account goes beyond the standard, Humean account of action determination. We already encountered the Humean desire-belief model above when I introduced intentions as a distinctive kind of mental state. Here is how Holton describes it:

> "All intentional action is explained just in terms of the agent's beliefs and desires. Agents act on whichever of their desires are strongest. An explanation of how agents stick by their resolutions must show how they thereby act on their strongest desires. (Insofar as resolutions are accepted as mental states at all, they must be thus reducible to beliefs and desires.)"[337]

[335] Richard Holton (2009), 112.
[336] Richard Holton (2009), 113.
[337] Richard Holton (2009), 112 f.

The Humean model tries to account for every action solely in terms of beliefs and desires. Hence if we want to explain an action that shows strength of will, we need to find the underlying beliefs and desires. Holton's suggestion enriches this model by adding willpower as a distinct faculty that is not reducible to desires and beliefs. With willpower as an additional determinant of actions, an action can express strength of will without being driven by the strongest desire.

A first consideration that speaks against the Humean model and in favor of the willpower model is that the desire-belief model "completely misrepresent[s] the phenomenology of the exercise of strength of will."[338] Typical cases in which we view people as exercising strength of will involve some sort of internal conflict. The agent has conflicting motivations and it is not easy to follow the course of action that is backed up by her judgment.

> "If these [desire-belief] accounts were right, then sticking to a resolution would consist in the triumph of one desire (the stronger) over another. But that isn't what it feels like. It typically feels as though there is a *struggle*. One maintains one's resolution by dint of effort in the face of contrary desire. […] by and large, maintaining strength of will requires effort."[339]

The problem of an implausible phenomenology arises because, in Humean accounts, the agent is always doing what she desires most. There is no room for action against one's desires. Of course, an agent might have conflicting desires, but since she always does what she desires most, she should never feel that she needs to push herself in a certain direction. But this is exactly what we feel sometimes: when we try to resist the next cigarette, the chocolate cake, or the TV, we regularly feel a severe inner conflict. The willpower model can make sense of this.

The starting point is the distinction between resolutions and desires, "for only then can we make sense of the idea of struggle involved in sticking with a resolution rather than bending to a desire."[340] The internal conflict that Holton envisages here

[338] Richard Holton (2009), 118.
[339] Richard Holton (2009), 118.
[340] Richard Holton (2009), 119.

has its source in the possibility of having a resolution that, when the time of action arrives, is opposed by strong desires. As we have seen, a resolution can persist in the face of contrary desires. In fact, that is the very rationale for forming resolutions. "Resolutions are contrary inclination defeating intentions: intentions formed by the agent with the very role of defeating any contrary inclinations that might emerge."[341] Hence, if we accept the idea that resolutions are not reducible to ordinary desires and beliefs, we are equipped with the right components to introduce real inner conflict.

Holton claims that "willpower is something that the agent actively employs."[342] It is an ability or capacity of the agent. As such, we can explain why "exercising willpower takes effort."[343] That is, we can explain the phenomenology of exercising willpower, namely, that "[i]t rather feels as though one is actively doing something, something that requires effort."[344] Using a capacity often is an effortful business. And in the case of willpower, we actually feel that we need to summon up energy in order to successfully resist a tempting influence. Holton gives an analysis of the effort involved in using willpower:

> "It is the mental effort of maintaining one's resolutions that is, of refusing to revise them. And my suggestion here is that one achieves this primarily by refusing to *reconsider* one's resolutions. On this picture, then, the effort involved in employing willpower is the effort involved in refusing to reconsider one's resolutions."[345]

It is plausible to assume that, given that willpower is an ability or capacity that the agent can actively employ, the effort for using it is a mental effort in contrast to a physical effort. The analysis of this effort in terms of refusing to reconsider one's resolutions is less obvious. Let me explain this.

Holton distinguishes between reconsideration and revision of a resolution. To revise a resolution is to let it drop and form an alternative one. Reconsideration, in

[341] Richard Holton (2009), 119.
[342] Richard Holton (2009), 120.
[343] Richard Holton (2009), 121.
[344] Richard Holton (2009), 121.
[345] Richard Holton (2009), 121.

contrast, is open to either a revision or a keeping of the resolution. If an agent reconsiders her intention, she contemplates the reasons that speak for and against it. At first glance, this kind of reconsideration appears to be no problem because deliberating about what one ought to do is, in general, a valuable procedure. But there lies a danger in reconsideration. "To fully reconsider a resolution is to open oneself to the possibility of revising it if the considerations come out a certain way; and that is to withdraw one's current commitment to it."[346] The problem lies in the aforementioned danger of a judgment shift, that is, of a deliberation that is shaped so that it fits to the strongest desire. "Although to reconsider a resolution is not, *ipso facto*, to revise it, it can be hard to keep the two separate. For, when temptation is great, its force will quickly turn a reconsideration into a revision."[347] For this reason, Holton views willpower as the ability to resist a reconsideration of one's resolution when one is faced with temptation.

Holton clarifies further that willpower is not needed anymore when certain behavior becomes automatic. The agent who automatically brushes her teeth before going to bed does not need to exercise willpower. She just automatically does it. Note that in this case, we don't have the feeling of an inner conflict. Willpower lies in the middle between automatic behavior and reconsideration, that is, suspension of a resolution. "What we need is a state that involves awareness of the resolution, and perhaps the considerations for which it is held, but which does not involve reconsideration. […] We thus need a state of awareness that falls short of suspension: what I shall call *rehearsal*."[348] Holton hastens to add that rehearsal and reconsideration lie on a continuum. Hence it is possible to slip from rehearsal to reconsideration without intending to do so. This is hardly a surprise because in rehearsing a resolution, we start to think about it and its merits, and we might be unable to suppress further thoughts about it.

Holton acknowledges this and mends his account of willpower accordingly. "It might be impossible to control whether we entertain the thought of having a cigarette. But it might be possible to control whether or not we go through the procedure that is involved in revising one's resolution not to."[349] The idea is that we have control over

[346] Richard Holton (2009), 121.
[347] Richard Holton (2009), 122.
[348] Richard Holton (2009), 123.
[349] Richard Holton (2009), 125.

our deliberation processes. In particular, we can control whether or not we deliberate a certain matter. The automaticity with which certain thoughts occur usually does not extend over whole deliberative processes. In other words, although the perception of a tempting alternative might automatically trigger a thought about how pleasant it would be to give in to the temptation, we are able to control whether we start to reconsider our resolution at this point. Willpower is the ability to block any reconsideration of resolutions.

What else apart from phenomenological support speaks in favor of the willpower account? We have empirical evidence that supports the willpower account.[350] Holton presents psychological studies of Walter Mischel who made a delay of gratification study with children. The children could get a reward immediately or they could wait, in which case they would get an even greater reward later. One of the findings was that children who employed strategies that helped them to avoid reconsideration were most successful in waiting for the larger reward. And it proved to be especially hard to delay the gratification if the smaller reward, let's say a cookie, is right in front of the children. A plausible interpretation is that seeing the cookie increases the pressure to reconsider one's resolution to wait for the larger reward.

Another source of evidence for the willpower account, and especially for the claim that willpower is a faculty, comes from studies which show that an agent's degree of strength of will is heavily influenced by such things as mood and emotional state, which are not themselves desires or resolutions: "the ability to abide by a resolution is affected by features that do not themselves seem to be desires or resolutions."[351] Holton refers to work of Baumeister, Heatherton and Tice[352] who showed that "[r]eformed alcoholics are far more likely to relapse if they are depressed, or anxious or tired."[353] This supports the general claim that actions are not only determined by desires, beliefs, and intentions. Hence, it casts doubts on the Humean model. With respect to strength of will, Holton goes on by pointing out that "states such as these [depression, anxiety, tiredness] affect one's ability to abide by *all* of one's resolutions: resolutions not to drink, not to smoke, to eat well, to exercise, to

[350] Walter Mischel (1996): 'From Good Intentions to Willpower', in: Peter Gollwitzer/John Bargh (1996) (eds.): *The Psychology of Action* (New York: Guilford Press), 197-218.
[351] Richard Holton (2009), 128.
[352] Roy Baumeister/Todd Heatherton/Diane Tice (1994): *Losing Control* (San Diego: Academic Press).
[353] Richard Holton (2009), 128.

work hard, not to watch daytime television or whatever."[354] The most plausible explanation for this is that the agent's ability to exert willpower is diminished by these emotional states. A Humean explanation that supposes that the agent's desires are collectively modified appears, in contrast, far-fetched.

Another phenomenon that speaks in favor of the claim that willpower is a faculty is so-called "ego depletion."[355] Ego depletion refers to a decrease in willpower due to tiring. "It appears that willpower comes in limited amounts that can be used up."[356] There is plenty of psychological evidence for this phenomenon. Baumeister, Bratslavsky, Muravan and Tice, for example, investigated the effect exercises of self-control have for subsequent tasks.[357] What they found is that agents who put an effort into eating radish instead of chocolate were less persistent on subsequent tasks, in this case solving a puzzle. In a similar study, subjects had to suppress their emotional responses to a movie, which diminished their persistence in holding a handgrip exerciser.[358] Holton continues by explaining that the suppression of emotional responses also "has an effect on resolutions: dieters eat more when they have been asked to suppress their emotional responses."[359] These observations support the willpower account. The best explanation for these observations is, "that one's action is determined not simply by the strength of one's desires and one's resolutions, but also by one's willpower; and that it is this component that is being affected by the repeated exercise."[360]

The last finding that Holton conjures up in support of the claim that willpower is a distinct faculty is, "that one can apparently develop one's faculty of willpower by repeated exercise."[361] That is, by exercising willpower, an agent becomes better and better at it. "Some research suggests that this might be right: subjects who undergo a regime of self-regulatory exercises – working on improving their posture for example – show markedly less tendency to suffer ego-depletion."[362] Again, the best explanation for this training effect seems to be that we have a distinct faculty of

[354] Richard Holton (2009), 128.
[355] Richard Holton (2009), 128.
[356] Richard Holton (2009), 128.
[357] Roy Baumeister/Ellen Bratslavsky/Mark Muravan/Diane Tice (1998): 'Ego-depletion: Is the Active Self a Limited Resource?', in: *Journal of Personality and Social Psychology* (74), 1252-1265.
[358] Compare Richard Holton (2009), 128.
[359] Richard Holton (2009), 129.
[360] Richard Holton (2009), 129.
[361] Richard Holton (2009), 129.
[362] Richard Holton (2009), 129.

willpower. The alternative explanation, according to which exercises of willpower in one case somehow systematically affect the strength of our desires or resolutions across the board, is highly implausible.

To sum this up: I follow Holton in assuming that willpower is a faculty in its own right. Agents possess more or less willpower. Strength of will consists in using one's willpower to uphold and execute one's intentions if difficulties occur. Before I explain more thoroughly how Holton's framework allows us to account for resolute agency, let me summarize the most important thoughts. First, agents can commit themselves strongly to a certain course of action by forming resolutions. Resolutions are intentions that commit the agent to continue with a certain course of action, even if the agent encounters detrimental influences. To say it with Holton, resolutions shield the agent against contrary inclinations. Second, weakness of will, which is, as I have highlighted, a threat to autonomy, consists in an over-readiness to drop one's intentions. That is, the weak-willed agent changes her intentions too easily. In other words, her commitment is very low. The strong-willed agent, in contrast, retains her intentions, even if this proves to be a difficult and effortful endeavor. Third, strength of will is partly grounded in the agent's willpower, whereby willpower is understood as a faculty sui generis. Agents use their willpower to retain and execute their intentions when they encounter obstacles and opposition.

To this point, I have reconstructed Holton's picture of strong willed and weak willed agency, in which the idea of a resolution plays a pivotal role. Against this background, we can develop a systematical understanding of resolute agency. I characterized resolute agency, on a general level, as agency that persists in conflicts and manages to reach goals in difficult situations. I distinguished two dimensions of resolute agency, namely, persistence and courage. For autonomy, it is of particular importance that the resolute agent is able to resist and overcome social pressure. I think that Holton's idea of resolutions and of the agent's dispositions to retain or to revise her resolutions provide the key to an adequate explication of resolute agency. What we need to do is extend Holton's framework so that it is broad enough to be applicable not only to cases of temptation, narrowly conceived, but to all kinds of conflicts. Holton himself points us in the direction in which his model can be extended. He mentions two general considerations. "First, not all resolutions will be explicit resolutions. Some will be more like general policies not to be moved in

certain ways."[363] This is an important aspect of resolute agency because the resolute agent is characterized primarily by certain policies that concern reasons for reconsidering her intentions.

> "Second, not all temptations will be what we would normally think of as temptations. Fear can knock one from a resolution; so can disgust. Again the process is much as before. Successful resistance requires a prior commitment, and the ability to stick to it achieved via non-reconsideration."[364]

At this point, Holton suggests using the notion of temptation as a technical notion that refers to all sorts of detrimental influences on one's resolve. Basically every influence that is able to undermine one's resolutions by leading to a reconsideration and subsequent revision of them can be deemed a temptation. Applied to persistence, this framework gives us the following characterization. Persistence consists in resolutions that commit the agent to retaining an intention, even if she is confronted with the need to try harder, try again, or try differently. A resolution is a complex intention or a set of two intentions. In addition to intending to perform a certain action, the agent who is resolved on a certain course of action also intends to retain this intention and to execute it, even if she encounters certain difficulties or conflicting motivations. By shielding herself against anticipated detrimental influences, the agent strengthens her commitment towards reaching her goal. If the agent is confronted with an obstacle, being resolved makes it more likely that she tries harder and, if necessary, adjusts her behavior because a resolution is meant to hold the agent on course even if it gets difficult. Thus, what I describe as persistence can be realized by the formation of resolutions. Resolutions are a kind of mental state that constitutes persistence.

We can also formulate this in terms of intention reconsideration. The persistent agent does not reconsider her intention when confronted with obstacles. Instead, she focuses her deliberation on how to overcome the obstacle. I use the notion of an obstacle in a very wide sense here. An obstacle can be a closed door, a canceled flight, or a plan that turns out to be inadequate for reaching the goal. The persistent agent is

[363] Richard Holton (2009), 135.
[364] Richard Holton (2009), 135.

not taken aback by such obstacles; instead, she searches for alternative courses of action that allow her to reach her goals.

With respect to autonomy, social pressure and how an agent reacts to it is of special significance. Holton's account of temptations and resisting temptation can be applied to this issue by saying that the autonomous agent is disposed to resist reconsideration of her commitment towards a certain action or project when she encounters social pressure against it. This disposition – courage, as I call it – is central to autonomy. The courageous agent forms resolutions with the aim of not being intimidated or led astray by social pressure. Hence, when she encounters social pressure, her resolutions shield her from reconsidering her intentions. She possesses enough willpower not to reconsider her intentions just because other people expect her to behave differently and threaten her with negative sanctions.

To summarize this: Holton's framework of strong willed agency is especially useful for explicating the dimension of autonomy that is concerned with inner strength and assertiveness, that is, resoluteness. The notion of a resolution as well as the account of willpower provides us with the conceptual resources to account for resolute agency. An agent is resolute if she is able to retain her intentions. And the retention of intentions is, to a large degree, a matter of forming strong resolutions and exercising willpower.

7.6 Resolute Agency and the Paradigmatic Cases of Non-autonomy

Throughout the discussion, I have presented some paradigmatic cases of non-autonomy, namely, compulsion, coercion, and manipulation. I used the intuition that these are indeed paradigmatic cases of non-autonomy in order to further describe the phenomenon that I am interested in when I talk about autonomous agency. That is, I used these examples to illuminate the concept of autonomy under discussion. I also said that an adequate account of this concept of autonomy ought to be able to explain why compulsion, coercion, and manipulation result in a loss of autonomy.

In Chapter 5, I discussed contemporary accounts of personal autonomy which put the idea of self-directedness in the center. It turned out that self-directedness accounts of autonomy are successful in explaining why compulsion issues in a loss of autonomy. Indeed, the contrast between compulsory agency and autonomous agency

is one of the main explanatory targets for these approaches, because a big part of this discussion takes its cues from Frankfurt's discussion of the unwilling addict. But as I argued above with regard to the cases of coercion and manipulation, the self-directedness accounts have a fundamental problem. It appears perfectly fine to imagine that an agent is self-directed although she is the victim of coercion or manipulation. At least the self-directedness accounts I discussed are compatible with this possibility. And the protagonists of this debate are well aware of this shortcoming. Bratman and Watson, for example, emphasize that their theories are incomplete when it comes to an adequate treatment of manipulation.

In reaction to some of the inadequacies that are attached to an understanding of autonomy under discussion solely in terms of self-directedness, I argued that we ought to acknowledge resolute agency as a distinctive feature of autonomy. At this point, the question arises whether the account of autonomy as resolute agency can deepen our understanding of these cases of non-autonomy.

The self-directedness accounts I discussed explain the fact that compulsion undermines autonomy by highlighting the difference between a will the agent identifies with and a will that is in some sense alien to her. Of course, the details of these accounts are open to different kinds of criticism and doubt. But in principle, they provide a reasonable answer to the problem of compulsion. When we agree that there is a reasonable way to account for the self vs. non-self distinction, then we can explain the threat compulsion poses for autonomy by pointing out that compulsive impulses come from the non-self. The problem with respect to coercion and manipulation is that the analogous argumentation does not hold. It is not plausible that an agent who gives in to a threat or who is the victim of manipulation is necessarily guided by impulses that have their source in the non-self. Let me clarify this issue a little bit further.

Different theories of self-directedness suggest different criteria for making the self vs. non-self distinction. What they all have in common is that they suggest a formal criterion. According to this formal criterion, being self-directed consists in possessing a certain psychological structure, for example, wholeheartedly endorsing a first-order desire or intending in accordance with one's evaluational standpoint. I call these criteria formal ones because, with respect to the content of these psychological states, the self-directedness accounts pose no constraints. For example, in Frankfurt's

hierarchical account, every first-order desire that is wholeheartedly endorsed, no matter what content, belongs to the agent's self, and the agent counts as self-directed in having this desire and acting on it. Similarly, in Watson's evaluational account, every intention that is backed up by the agent's evaluational standpoint constitutes the agent's autonomous will. Neither the content of the intention nor the content of the agent's reasons matter as long as they fit. And this is exactly the reason why self-directedness accounts have a problem with explaining why coercion and manipulation undermine autonomy. It is possible that an agent desires to give in to a threat or intends to give in to a threat because she judges that this is what she ought to do. In these cases she would count as self-directed because she fulfills the formal criterion. In an analogous way, she would count as self-directed even if she meets the criterion as a consequence of manipulation.

Against the background of an understanding of autonomy that also includes resolute agency, there appears to be a way to solve this problem – at least in outline. The reason is, as we will see, that the notion of resolute agency has a substantial dimension, that is, it does not only describe a certain structure of mental states or certain mechanisms regardless of their content. Quite to the contrary, it refers to dispositions and abilities to react in specific ways to specific kinds of influences. Let me first discuss the case of coercion and then contemplate the case of manipulation.

It appears that the case of coercion can be dealt with given that we understand autonomy in terms of resolute agency. The resolute agent is disposed to make her own decisions without being intimidated by other people's demands. I developed this idea at some length, referring to it in a technical sense as courage. Autonomy is partly constituted by courage. Coercion directly attacks an agent's courage. It aims at intimidating its victim. Someone who uses coercion wants to subjugate another person's will by invoking fear. The victim of coercion is confronted with additional negative consequences for certain courses of action. The person who coerces another tries to determine her behavior by threatening her. Courage is the character trait that directly opposes coercive attempts. Hence, if we acknowledge that resolute agency is characterized by a disposition not to give in to threats, we can explain why coercion undermines autonomy. The reason is that the aim of coercion is to let the victim subjugate her will because of fear.

One might object that this explanation is somewhat circular. We start with the intuition that coercion violates autonomy. Based on this intuition, we develop an account of autonomy in which the notion of resolute agency plays an essential role. And finally, we use this account to explain the intuition. It is hardly surprising that coercion can be explained as a threat to autonomy when the account we use for this explanation has been modeled on the assumption that coercion undermines autonomy.

I think that this objection is fair as far as it goes. It is correct to point out that I introduced the notion of resolute agency as a dimension of autonomy because I assumed that giving in to threats, being intimidated into submission, and other such things undermine autonomy. I did not develop this idea independently, and I applied it subsequently to the case of coercion. But this does not imply that the notion of resolute agency has no explanatory value with respect to cases of coercion.

I emphasized right from the start that I use these paradigmatic cases of non-autonomy as a starting point for developing an account of autonomy. The aim in this context is not to convince a person who doubts that these cases undermine autonomy. If someone thinks that autonomy is compatible with compulsion, coercion, or manipulation, this person has a different concept in mind when she talks about autonomy. Based on the assumption that compulsion, coercion, and manipulation undermine autonomy, we can search for an understanding of autonomy that makes sense of this. And this is what I did when I started to account for personal autonomy in terms of resolute agency. The notion of resolute agency does not just describe certain formal characteristics of agency. It is a substantial one insofar as it uses notions such as intimidation.

The understanding of autonomy as resolute agency also points in the direction of an explanation why manipulation undermines autonomy. Manipulation is an external interference with the agent's intention formation and her self-creation. It is a hallmark of resolute agency that the resolute agent is able to resist any interference with her intention formation. I explicated this resistance in terms of resolution formation and the possession of willpower. Manipulation circumvents the agent's resoluteness. The agent cannot use her willpower because the manipulative attempt meddles with her standpoint without the agent's awareness. The effect is that an agent's resoluteness is rendered ineffectual. And this undermines the agent's

autonomy because resoluteness is an important dimension of autonomy. Another agent determines part of her practical identity and intention formation.

7.7 Conclusion

In this chapter, I have explicated the notion of resolute agency as an essential dimension of autonomy. Resoluteness consists in abilities and dispositions to prevail in situations of conflict. The resolute agent is able to retain her intentions against opposition. I distinguished two dimensions of resolute agency, namely, persistence and courage. Persistence is concerned with the agent's dispositions and abilities to try hard, try again, or try differently, if this becomes necessary. Courage refers to those dispositions and abilities that allow the agent to resist social pressure. I argued that some aspects of the concept of autonomy under consideration are best explained in terms of resolute agency. Holton's account of resolutions and willpower proved to be especially useful to explicate the two dimensions of resolute agency in greater depth. An account of resolute agency is an important contribution to an explication of the concept of autonomy for two reasons. First, dispositional autonomy is best described in terms of resolute agency. Second, local autonomy does not only consist in self-directed agency, but also in resolute agency.

This concludes my discussion. In the final chapter, I summarize the most important results.

8. Self-Directedness and Resoluteness. The Two Dimensions of Autonomy

I started this study with the example of Antigone – the very first person we know of who was ever called autonomous. The aim of my discussion was to explore and explicate the concept of autonomy that is exemplified by such agents as Antigone, Marie Curie, and Martin Luther King Jr. The core of this concept is formed by the idea that the autonomous agent is able to shape her life against opposition and in accordance with her own desires, beliefs, and values. This understanding of autonomy has two dimensions: self-directedness and resoluteness. Autonomy consists in abilities and dispositions to develop one's own authentic standpoint and to express it in one's life.

In Chapter 1, I presented the historical origins of the concept of autonomy under investigation. The aptitude to prevail in conflicts was a central feature of the original political notion of autonomy as it was developed in ancient Greece. This aptitude, conceived of in a broad sense, is what I call resoluteness in the context of personal autonomy. Kant opened the door for an understanding of autonomy as a property of persons. And Frankfurt put the idea of being true to one's authentic character in the center of the autonomy debate – an idea to which I refer with the notion of self-directedness.

Chapter 2 delineated the kind of concept that is the target of the discussion in more systematical detail. According to this concept, autonomy is a natural and gradual property of persons. In particular, it is a set of dispositions and abilities that enable the agent to develop her own authentic standpoint and to shape her life accordingly against opposition. Finally, autonomy is not tied to moral norms and it does not presuppose alternative possibilities, that is, autonomy differs from free will.

I discussed in Chapter 3 what I mean by speaking of autonomy as a natural property. According to this idea of natural autonomy, autonomous agency is ontologically explicable within a naturalistic framework. Despite contrary intuitions, according to which autonomy requires an agent to transcend nature, the concept of autonomy under discussion conceives of autonomous agency within the limits of nature. Dualistic approaches, recourse to agent causation, or the introduction of volitions contradict this idea. The notion of natural autonomy has to be explicated within an event-causal framework.

In Chapter 4, I developed the action theoretical foundation on which my understanding of autonomous agency is built. I presented a naturalistic account of agency and action, according to which an action is behavior that is caused in the right way by the agent's intentions. In order to corroborate this claim, I started with a characterization of actions as exercises of control. I then argued that the right kind of control is realized by a correct functioning of the agent's intentions.

Chapter 5 discussed the notion of self-directedness. Autonomous agents manage to do what they *really* want to do instead of just following another person's lead or succumbing to motives that violate their own standpoint. An account of self-directedness explains what it means that an agent's will is in this emphatic sense her own. I argued that self-directedness is an essential dimension of autonomy and discussed the most important approaches towards an understanding of self-directed agency. It turned out that an agent's evaluational standpoint provides the criterion for deciding what self-directedness consists in for a particular agent. Self-directedness consists in being true to one's practical identity.

I pointed out that self-directedness accounts of autonomy dominate the autonomy debate. But they fall short of giving a complete explication of the concept of autonomy that I highlighted before. In particular, they lack the resources to account for the antagonistic dimension of autonomy. An agent's aptitude to prevail in conflicts lies beyond the scope of an account of self-directedness. I pointed this out by looking at exemplary cases of non-autonomy, in particular, coercion and manipulation.

In Chapter 6, I presented considerations that further demonstrate why the concept of autonomy under consideration cannot be captured solely in terms of self-directedness. The second dimension of autonomy that I dubbed resolute agency is not reducible to self-directed agency. Examples of highly resolute but insufficient self-directed agents, as well as examples of agents who are highly self-directed but lack resoluteness, back up my claim that autonomous agency is constituted both by self-directedness as well as resoluteness.

The final chapter, Chapter 7, investigated resolute agency. I developed an account of resolute agency around the key notions of persistence and courage. Persistence is a set of abilities and dispositions to try hard, to try harder, and to try differently if obstacles occur. Courage is a set of dispositions and abilities to resist

social pressure. Both aspects of resoluteness are constitutive of autonomy. The autonomous agent is able to shape her life in detrimental circumstances, that is, she is able to overcome obstacles and resist pressure in order to shape her life as an authentic expression of her practical identity. I presented Holton's account of resolutions and willpower in order to explicate the notion of resolute agency in more detail.

*

The autonomy debate since Frankfurt has focused on getting a handle on the notion of self-directedness. As I have argued, self-directed agency is indeed an essential aspect of autonomous agency. We think of agents as being autonomous partly because they are able to develop and express *their own* standpoint instead of doing what others expect them to do. The value that is associated with self-directedness is the value of being authentic. However, if we try to explicate autonomy solely in terms of self-directedness, we lose sight of a dimension of autonomy that not only importantly shapes our intuitive understanding of autonomy, but also turns out to be the source of another special value attached to autonomy. The autonomous agent is a strong agent that manages to express what is important to her in situations where this proves to be hard because of obstacles and opposition she encounters. When we say that autonomy calls for our respect, we have this dimension of autonomy in mind because autonomy cannot be taken for granted. It is a personal achievement. In order to come to terms with this dimension of autonomy, I introduced and explicated the notion of resolute agency. This notion captures the idea that autonomous agents are characterized by abilities and dispositions to stand their ground in the face of difficulties and opposition.

The notion of resolute agency is not only essential for explicating the idea that autonomous agents possess significant strength and assertiveness in going after their goals. In addition to that, it also turns out to be the key to an adequate understanding of autonomy as a dispositional property of persons. Self-directedness is a local property, whereas resoluteness is a dispositional property. According to the concept of autonomy that I explored in this study, we typically characterize persons as autonomous because they display certain dispositions and abilities in the way they shape their whole life. Often, when we contemplate a person's autonomy, we are not so much concerned with one particular situation, but with the person's whole pattern

of making decisions and realizing her projects. We think of the autonomous person as the person who *will* behave in certain ways in the future or who *would* behave in certain ways in some counterfactual scenario. The foundation for these kinds of judgments is a dispositional understanding of autonomy. And as I argued, for an explication of this dispositional understanding we cannot rely on an account of self-directedness, but need to refer to an account of resoluteness instead.

The account of resoluteness that I presented perfectly fits the action theoretical background that I have developed. Persistence and courage, the two dimensions of resoluteness, are constituted by dispositions and abilities that concern the formation and execution of intentions. Intentions are the mental state that realizes the kind of agential control required for action. The autonomous agent is able to make a special use of her agential control, that is, her abilities and dispositions to form and execute intentions. A characteristic of autonomous agents is that they can form intentions that concern the ways in which they form other intentions. They exert hierarchical control. The account of resolute agency explicates these dispositions and abilities in terms of particular kinds of intentions, namely, self-governing policies, implementation intentions, and resolutions, and a special capacity that agents can use in order to retain their intentions, namely, willpower. By forming these kinds of intentions, and by using their willpower, agents resolutely go after their goals.

The conceptual distinction of resolute agency and self-directed agency is important in order to delineate the different dimensions of autonomy. But we should not overlook the actual interplay of these two dimensions in autonomous agency. As a matter of fact, achieving self-directedness requires the agent to be resolute. Only being resolute enables an agent to be self-directed. And this connection also influences the way we conceive of resoluteness. We value resoluteness not least because it plays this significant role in achieving self-directedness. For a full understanding of autonomy, we need to acknowledge the importance of both of these dimensions.

This concludes my discussion. The central insight of this study is that resolute agency is an essential aspect of autonomy. Dispositional autonomy is constituted by an agent's dispositions and abilities to overcome obstacles. The autonomy debate underemphasized this aspect of autonomy by solely focusing on self-directed agency. But as important as self-directedness is for autonomous agency, without resoluteness,

an agent falls short of being autonomous. The personal strength that resoluteness consists in is a hallmark of autonomous agents like Antigone, Marie Curie, and Martin Luther King Jr. This study attempts to put it back into the center of our struggles to come to terms with the idea of autonomy.

Bibliography

Aguilar, Jesus H./Buckareff, Andrei A. (2009): 'Agency, Consciousness, and Executive Control', in: *Philosophia* (37), 21-30.

Anderson, Joel/Honneth, Axel (2005): 'Autonomy, Vulnurability, Recognition, and Justice', in: John Christman/Joel Anderson (eds.) (2005), 127- 149.

Aristotle (1985): *Nicomachean Ethics*, trans. by Terence Irwin (Indianapolis: Hackett Publishing).

Armstrong, D. M. (1997): *A world of states of affairs* (Cambridge: Cambridge University Press).

Asch, Solomon (1956): 'Studies of Independence and Conformity. A Minority of one against an unanimous majority', in: *Psychological Monographs* (70), 1-70.

Bainton, Roland H. (1950): *Here I Stand. A Life of Martin Luther* (Nashville: Abingdon Press).

Baumeister, Roy/Heatherton, Todd/Tice, Diane (1994): *Losing Control* (San Diego: Academic Press).

Baumeister, Roy/Bratslavsky, Ellen/Muravan, Mark/Tice, Diane (1998): 'Ego-depletion: Is the Active Self a Limited Resource?', in: *Journal of Personality and Social Psychology* (74), 1252-1265.

Baumeister, Roy F./Masicampo, E. J./Vohs , Kathleen D. (2011): 'Do Conscious Thoughts Cause Behavior?', in: *Annual Review of Psychology* (62), 331-361.

Bayne, Tim: 'The Sense of Agency', in: Fiona Macpherson (ed.) (2011), 355-374.

Benson, John (1983): 'Who is the Autonomous Man?', in: *Philosophy* (223), 5-17.

Berofsky, Bernard (1995): *Liberation from the Self. A Theory of Personal Autonomy* (Cambridge: Cambridge University Press).

Bishop, John (1989): *Natural Agency. An Essay on the Causal Theory of Action* (Cambridge: Cambridge University Press).

Brand, Myles (1984): *Intending and Acting. Toward a Naturalized Action Theory* (Cambridge MA: MIT Press).

Brass, Marcel/Haggard, Patrick (2008): 'The What, When, Whether Model of Intentional Action', in: *The Neuroscientist* (14), 319-325.

Bratman, Michael R. (1987): *Intention, Plans, and Practical Reason* (Cambridge MA: Harvard University Press).

Bratman, Michael R. (2007): *Structures of Agency. Essays* (Oxford: Oxford University Press).

Bratman, Michael R. (2007 a): 'Introduction', in: Michael R. Bratman (2007), 3-18.

Bratman, Michael R. (2007 b): 'Autonomy and Hierarchy', in: Michael R. Bratman (2007), 162-186.

Bratman, Michael, R. (2007 c): 'Planning Agency, Autonomous Agency', in: Michael R. Bratman (2007), 195-221.

Bratman, Miachel R. (2007 d): 'Three Theories of Self-Governance', in: Michael R. Bratman (2007), 222-253.

Broad, C. D. (1952): 'Determinism, Indeterminism, and Libertarianism', in: C. D. Broad (1952): *Ethics and the History of Philosophy* (London: Routledge and Kegan Paul), 195-217.

Buss, Sarah (1994): 'Autonomy Reconsidered', in: *Midwest Studies in Philosophy* (19), 95-121.

Chisholm, Roderick M. (1966): 'Freedom and Action', in: Keith Lehrer (ed.) (1966), 11-44.

Christman, John (1989) (ed.): *The Inner Citadel. Essays on Individual Autonomy* (Oxford: Oxford University Press).

Christman, John/Anderson, Joel (eds.) (2005): *Autonomy and the Challenges to Liberalism* (Cambridge: Cambridge University Press).

Damasio, Antonio R. (1994): *Descartes' Error. Emotion, Reason, and the Human Brain* (New York: G. P. Putnam's Sons).

Dancy, Jonathan (2000): *Practical Reality* (Oxford: Oxford University Press).

Darwall, Stephen (2006): 'The Value of Autonomy and Autonomy of the Will', in: *Ethics* (116), 263-284.

Davidson, Donald (1980): *Essays about Actions and Events* (Oxford: Oxford University Press).

Davidson, Donald (1980 a): 'Actions, Reasons, and Causes', in: Donald Davidson (1980), 3-19.

Davidson, Donald (1980 b): 'Freedom to Act', in: Donald Davidson (1980), 63-82.

Descartes, René (1641/1986): *Meditations on First Philosophy*, trans. by John Cottingham (Cambridge: Cambridge University Press).

Dretske, Fred (1988): *Explaining Behavior. Reasons in a World of Causes* (Cambridge, MA: MIT Press).

Dworkin, Gerald (1988): *The Theory and Practice of Autonomy* (Cambridge: Cambridge University Press).

Ekstrom, Laura Waddell (1993): 'A Coherence Theory of Autonomy', in: *Philosophy and Phenomenological Research* (53), 599-616.

Ekstrom, Laura Waddell (2005): 'Autonomy and Personal Integration', in: James Stacey Taylor (ed.) (2005), 143-161.

Enç, Berent (2003): *How We Act. Causes, Reasons, and Intentions* (Oxford: Oxford University Press).

Feinberg, Joel (1989): 'Autonomy', in: John Christman (ed.) (1989), 27-53.

Figueira, Thomas (1990): 'Autonomoi Kata Tas Spondas (Thucydides 1.67.2)', in: *Bulletin of the Institute of Classical Studies* (37), 63-88.

Firth, Chris D./Blakemore, Sarah-Jayne/Wolpert, Daniel M. (2000): 'Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action', in: *Brain Research Reviews* (31), 357-363.

Fisher, John Martin/Ravizza, Mark (1998): *Responsibility and Control. A Theory of Moral Responsibility* (Cambridge: Cambridge University Press).

Forst, Rainer (2005): 'Political Liberty: Integrating Five Conceptions of Autonomy', in: John Christman/Joel Anderson (eds.) (2005), 226-242.

Frankel Paul, Ellen/Miller, Jr., Fred D./Paul, Jeffrey (eds.) (2003): *Autonomy* (Cambridge: Cambridge University Press).

Frankfurt, Harry G. (1988): *The Importance of What We Care About* (Cambridge: Cambridge University Press).

Frankfurt, Harry G. (1988 a): 'Freedom of the Will and the Concept of a Person', in: Harry G. Frankfurt (1988), 11-25.

Frankfurt, Harry G. (1988 b): 'Identification and Externality', in: Harry G. Frankfurt (1988), 58-68.

Frankfurt, Harry G. (1988 c): ‘Identification and Wholeheartedness’, in: Harry G. Frankfurt (1988), 159-176.

Frankfurt, Harry G. (1999): *Necessity, Volition, and Love* (Cambridge: Cambridge University Press).

Frankfurt, Harry G. (1999 a): ‘The Faintest Passion’, in: Harry G. Frankfurt (1999), 95-107.

Frankfurt, Harry G. (1999 b): ‘Autonomy, Necessity, and Love’, in: Harry G. Frankfurt (1999), 129-141.

Frankfurt, Harry G. (2004): *The Reasons of Love* (Princeton: Princeton University Press).

Friedrich, Daniel (2011): *Autonomy and Subservience* (Manuscript).

Goldman, Alvin (1971): *A Theory of Human Action* (New Jersey: Prentice Hall).

Gollwitzer, Peter M. (1993): ‘Goal Achievement: The Role of Intentions’, in: *European Review of Social Psychology* (4), 141-185.

Gollwitzer, Peter M./Bargh, John (1996) (eds.): *The Psychology of Action* (New York: Guilford Press).

Gollwitzer, Peter M./Schaal, Bernd (1998): ‘Metacognition in Action: The Importance of Implementation Intentions’, in: *Personality and Social Psychology Review* (2), 124-136.

Gollwitzer, Peter M./Gawrilow, Caterina/Oettingen, Gabriele (2010): ‘The Power of Planning: Self-Control by Effective Goal-striving, in: R.R. Hassin/K.N. Ochsener/Y. Trope (eds.) (2010), 3-26.

Goschke, Thomas (2003): ‘Voluntary action and cognitive control from a cognitive neuroscience perspective’, in: Sabine Maasen/Wolfgang Prinz/Gerhard Roth (eds.) (2003), 49-85.

Guignon, Charles (2008): ‘Authenticity’, in: *Philosophy Compass* (3/2), 277-290.

Haggard, Patrick (2008): ‘Human volition: towards a neuroscience of will’, in: *Nature Reviews Neuroscience* (9), 934-946.

Haji, Ishtiyaque/Cuypers, Stefaan E. (2008): ‘Authenticity-Sensitive Preferentism and Educating for Well-Being and Autonomy’, in: *Journal of Philosophy of Education* (42), 85-106.

Hassin, R. R./Ochsener, K. N./Trope, Y. (eds.) (2010): *Self Control in Society, Mind, and Brain* (Oxford: Oxford University Press).

Heath, Joseph (2005): 'Liberal Autonomy and Consumers Sovereignty', in: John Christman/Joel Anderson (eds.) (2005), 204-225.

Heckhausen, Heinz/Gollwitzer, Peter (1987): 'Thought contents and cognitive functioning in motivational versus volitional states of mind', in: *Motivation and Emotion* (11), 101-120.

Holton, Richard (2009): *Willing, Wanting, Waiting* (Oxford: Oxford University Press).

Hyun, Insoo (2001): 'Authentic Values and Individual Autonomy', in: *The Journal of Value Inquiry* (35), 195-208.

Kane, Robert (1998): *The Significance of Free Will* (Oxford: Oxford University Press).

Kant, Immanuel (1785/1965): *Groundwork of the Metaphysic of Morals*, trans. by H. J. Paton (New York: Harper Perennial).

Korsgaard, Christine M. (1996): *The Sources of Normativity* (Cambridge: Cambridge University Press).

Kuhl, Julius (1985): 'Volitional Mediators of Cognition-Behavior Consistency: Self-Regulatory Processes and Action Versus State Orientation, in: Julius Kuhl/Jürgen Beckmann (eds.) (1985), 101-128.

Kuhl, Julius/Beckmann, Jürgen (eds.) (1985): *Action Control. From Cognition to Behavior* (Berlin: Springer-Verlag).

Kuhl, Julius/Beckmann, Jürgen (1985): 'Historical Perspectives in the Study of Action Control', in: Julius Kuhl/Jürgen Beckmann (eds.) (1985), 89-100.

Lehrer, Keith (ed.) (1966): *Freedom and Determinism* (New York: Random House).

Lehrer, Keith (2003): 'Reason and Autonomy', in: Ellen Frankel Paul/Fred D. Miller, J./Jeffrey Paul (eds.) (2003), 177-198.

Maasen, Sabine/Prinz, Wolfgang/Roth, Gerhard (eds.) (2003): *Voluntary Action. Brains, minds, and sociality* (Oxford: Oxford University Press).

Macpherson, Fiona (ed.) (2011): *The Senses. Classic and Contemporary Philosophical Perspectives* (Oxford: Oxford University Press).

McGinn, Colin (1982): *The Character of Mind* (Oxford: Oxford University Press).

Mele, Alfred R. (1992): *Springs of Action. Understanding Intentional Behavior* (Cambridge: Cambridge University Press).

Mele, Alfred R. (1995): *Autonomous Agents. From Self-Control to Autonomy* (Oxford: Oxford University Press).

Mele, Alfred R. (2009): *Effective Intentions. The Power of Conscious Will* (Oxford: Oxford University Press).

Miller, Earl K./Cohen, Jonathan D. (2001): 'An Integrative Theory of Prefrontal Cortex Function', in: *Annual Review of Neuroscience* (24), 167-202.

Millgram, Elijah (ed.) (2001): *Varieties of Practical Reasoning* (Cambridge MA: MIT Press).

Mischel, Walter (1996): 'From Good Intentions to Willpower', in: Peter Gollwitzer/John Bargh (1996) (eds.), 197-218.

Oshana, Marina A. L. (1998): 'Personal Autonomy and Society', in: *Journal of Social Philosophy* (29), 81-102.

Oshana, Marina A. L. (2002): 'The Misguided Marriage of Responsibility and Autonomy', in: *The Journal of Ethics* (6), 261-280.

Ostwald, Martin (1982): *Autonomia: Its Genesis and Early History* (Atlanta: Scholars Press).

Pacherie, Elisabeth (2000): 'The Content of Intentions', in: *Mind & Language* (15), 400-432.

Pacherie, Elisabeth (2008): 'The Phenomenology of Action: A conceptual framework', in: *Cognition* (107), 179-217.

Pacherie, Elisabeth/Haggard, Patrick (2010): 'What are Intentions?', in: Walter Sinnott-Armstrong/Lynn Nadel (2010): *Conscious Will and Responsibility* (Oxford: Oxford University Press), 70-84.

Papineau, David (1993): *Philosophical Naturalism* (Oxford: Blackwell Publishers).

Parfit, Derek/Broome, John (1997): 'Reasons and Motivation', in: *Proceedings of the Aristotelian Society. Supplementary Volumes* (71), 99-146.

Pauen, Michael (2004): *Illusion Freiheit? Mögliche und unmögliche Konsequenzen der Hirnforschung* (Frankfurt a. M.: S. Fischer Verlag).

Pauen, Michael (2008): *Autonomy* (Manuscript).

Pettit, Philipp (1987): 'Humeans, Anti-Humeans, and Motivation', in: *Mind* (96), 530-533.

Raz, Joseph (1986): *The Morality of Freedom* (Oxford: Oxford University Press).

Ryle, Gilbert (1949): *The Concept of Mind* (London: Hutchinson).

Sartre, Jean Paul (1948): *Existentialism and Humanism* (London: Methuen).

Scanlon, T. M. (2000): *What We Owe to Each Other* (Cambridge MA: Harvard University Press).

Schueler, G. F. (1995): *Desire. Its role in Practical Reason and the Explanation of Action* (Cambridge MA: MIT Press).

Schueler, G. F. (2003): *Reasons & Purposes. Human Rationality and the Teleological Explanation of Action* (Oxford: Oxford University Press).

Shiffrin, R. M./Schneider, W. (1977): 'Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory', in: *Psychological Review* (84), 127-190.

Smith, Adam (1982): *The Theory of Moral Sentiments*, (D.D. Raphael/ A.L. Macfie (eds.): Vol. I of the Glasgow Edition of the Works and Correspondence of Adam Smith; Indianapolis: Liberty Fund).

Smith, Michael (1987): 'The Humean Theory of Motivation', in: *Mind* (96), 36-61.

Snare, Francis (1991): *Morals, Motivation and Convention* (Cambridge: Cambridge University Press).

Sophocles: *Antigone*, trans. by William Blake Tyrrell/Larry J. Bennett (http://www.stoa.org/diotima/anthology/ant/antigstruct.htm).

Taylor, Charles (1985): *Human Agency and Language. Selected Papers 1* (Cambridge: Cambridge University Press).

Taylor, Charles (1985 a): 'What is human agency?', in: Charles Taylor (1985), 15-44.

Taylor, Charles (1989): *The Sources of the Self: The Making of the Modern Identity* (Cambridge: Cambridge University Press).

Taylor, James Stacey (ed.) (2005): *Personal Autonomy. New Essays on Personal Autonomy and Its Role in Contemporary Moral Philosophy* (Cambridge: Cambridge University Press).

Thagard, Paul (2001): 'How to Make Decisions: Coherence, Emotion, and Practical Inference', in: Elijah Millgram (ed.) (2001), 355-371.

Thucydides (2009): *The Peloponnesian War*, trans. by Martin Hammond (Oxford: Oxford University Press).

Van Inwagen, Peter (1983): *An Essay on Free Will* (Oxford: Clarendon Press).

Velleman, J. David (1992): 'What happens when someone acts?', in: *Mind* (101), 461-481.

Velleman, J. David (1996): 'The Possibility of Practical Reason', in: *Ethics* (106), 694-726.

Velleman, J. David (2006): 'The Self as a Narrator', in: J. David Velleman (2006): *Self to Self. Selected Essays* (Cambridge: Cambridge University Press), 203-223.

Wallace, Jay (1990): 'How to argue about practical reason', in: *Mind* (99), 355-385.

Walter, Henrik (2001): *Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy* (Cambridge MA: MIT Press).

Watson, Gary (1975): 'Free Agency', in: *Journal of Philosophy* (72), 205-220.

Watson, Gary (1987): 'Free Action and Free Will', in: *Mind* (96), 145-172.

Wegner, Daniel M. (2002): *The Illusion of Conscious Will* (Cambridge, MA: Harvard University Press).

Wolf, Susan (1990): *Freedom Within Reason* (Oxford: Oxford University Press).